The
University
Of
Sheffield.

# “It ain’t all good”
# Machinic abuse detection and marginalisation in machine learning

**By:**

Zeerak Mustafa Talat Khan

This dissertation is submitted for the degree of
*Doctor of Philosophy in Computer Science*

The University of Sheffield
Department of Computer Science
Faculty of Engineering

May 2021

*For Ammi,*
*Without your sacrifice, love and support*
*these pages would remain blank.*

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 80,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Zeerak Mustafa Talat Khan
May 2021

# Acknowledgements

The journey of undertaking this Ph.D. has been a labour of love and it is a journey that I would not have been able to complete without the friendship and support of peers and mentors.

First, I would like to thank my supervisor, Kalina Bontcheva, for her unending support, generosity, grace and invaluable advice throughout the entire process. I would also like to acknowledge Mehmet Bakir, my second supervisor for his advice and willingness to sit with me and work through technical challenges. Here, I would also like to thank Andreas Vlachos, who was briefly my second supervisor. Without his insistence and support, my technical knowledge and understanding would not be what it has become.

I would also like to take a moment to acknowledge the teachers and figures of authority throughout my childhood and youth who have found creative ways to discourage me from academic pursuits.
In contrast, I would like to thank Dirk Hovy, Joel Tetreault, Wendy Hui Kyong Chun, Barbara Plank, and Leon Derczynski for their advice for each talking me out of dropping out of the programme and encouraging me to flourish. I cannot overstate my thanks to the workshop attendees and organisers at the Lorentz workshop who created an environment that showed that even within academia's elitist hostility and contempt of marginalised bodies, there are breathing spaces and homes to be found.

Similarly so, I would also like to take a moment to thank my collaborators and co-conspirators, Nanna Bonde Thylstrup, Smarika Lulz, Isabelle Augenstein, James Thorne, and Joachim Bingel, Maya Indira Ganesh, Su Lin Blodgett, Dan Bareket, Anjalie Field, Bertie Vidgen, Aline Villavicencio, Joris van Hoboken, and Frencien Dechene whose friendship and endless willingness to share their knowledge and expertise is reflected in every word.

I also want to acknowledge the musicians Chaka Khan and De La Soul, whose musical masterpiece "All Good" I quote in the title.

Throughout the time I've been working towards writing this dissertation, I have been organising and hosting workshops on online abuse and diversifying ACL conferences and the Natural Language Processing field. I want to thank each of my co-organisers for the innumerable ways in which they've helped make me a more well rounded researcher. I would also like to take this moment to apologise and thank my friends Adina, Nanna, Kayla, Dalton, Paul, Amy, and Yanai who have all suffered through drafts and the tortured prose in earlier drafts.

Throughout the time I have been working towards this dissertation, I have enjoyed the company and discussions of some of the smartest, most compassionate and kind people. Although a complete list of everyone would fill up another 180 pages, I want to in particular thank Lianna, Rasmus, Tal, Hector, Martin, Kasper, Søren, Noam, Asbjørn, Carl Johan, Nikita, Mareike, Hannah, Abeba, Deb, Alex, Alison, Haiyang, Fred, Ryan, Teresa, Libby, Amittai, George, Hardy, Valentina, Suchin, Amandalynne, Chiraag, Megan, Vinod, Jacque, Nikos, Swabha, Eleanor, Pranava, and Teck and Michael who have helped, discussed, laughed, and argued with me, always just at the right time.

I would like to thank my mother, Talat, and sisters, Rushna, Tehneyat and Aqleema, who have supported me and helped shaped my thinking. I also want to thank my niece Sofia and my nephews, Aiden, Miles, and Lukas, whose love warms my bones and are my reasons to work on this topic. Finally, I would like to acknowledge my partner, Fern to whom no amount of thanks can do justice in the ways she has helped me through this.

# Abstract

Online abusive language has been given increasing prominence as a societal problem over the past few years as people are increasingly communicating on online platforms. This increase in prominence has resulted in an increase in academic attention to the issue, particularly within the field of Natural Language Processing (NLP), which has proposed multiple datasets and machine learning methods for the detection of text-based abuse. Recently, the issue of disparate impacts of machine learning has been given attention, showing that marginalised groups in society are disproportionately negatively affected by automated content moderation systems. Moreover, a number of challenges have been identified for abusive language detection technologies, including poor model performance across datasets and a lack of ability of models to contextualise potentially abusive speech within the context of speaker intentions. This dissertation aims to ask how NLP models for online abuse detection can address issues of generalisation and context.

Through critically examining the task of online abuse detection, I highlight how content moderation acts as protective filter that seeks to maintain a sanitised environment. I find that when considering automated content moderation systems through this lens, it is made clear that such systems are centred around experiences of some bodies at the expense of others, often those who are already marginalised.

In efforts to address this, I propose two different modelling processes that a) centre the the mental and emotional states of the speaker by representing documents through the Linguistic Inquiry and Word Count (LIWC) categories that they invoke, and using Multi-Task Learning (MTL) to model abuse, such that the model takes aims to take account the intentions of the speaker.

I find that through the use of LIWC for representing documents, machine learning models for online abuse detection can see improvements in classification scores on in-domain and out-of-domain datasets. Similarly, I show that through a use of MTL, machine learning models can gain improvements by using a variety of auxiliary tasks that combine data for content moderation systems and data for related tasks such as sarcasm detection.

Finally, I critique the machine learning pipeline in an effort to identify paths forward that can bring into focus the people who are excluded and are likely to experience harms from machine learning models for content moderation.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

As people consume, interact with, and create online media at an ever growing rate, it becomes increasingly important to carefully consider the content with which they interact, how this influences and impacts audiences and the cultures that are formed in online spaces. Online abuse is a small but prominent portion of online communication, that is expressed through various forms , including cyber-bullying, hate speech, harassment, aggressive speech, and offensive speech (Vidgen et al., 2019). Extended exposure to such content can have adverse effects for users' engagement in online spaces (Fisher and McBride, 2016), lead to psychological harms for the targets of abuse (Gelber and McNamara, 2016), and be a factor in increases in hate crimes (Müller and Schwarz, 2020). As a result, social media platforms have long been subject to pressure to moderate and remove such content from users and regulators alike. Moderating, removing, and adjudicating content in online spaces has traditionally been a human effort (Roberts, 2019) however as computational methods such as machine learning have matured, social media platforms such as Facebook have increasingly come to rely on such automated methods (Facebook, n.d.). . These automated content moderation systems developed by commercial platforms operate across several different modalities, including images, text, and videos. Along with this increase in automated content moderation systems developed by commercial entities, there has been an increase in attention to the challenge of developing automated systems for detecting text-based abuse in the Natural Language Processing (NLP) community. Beyond the NLP community, there has been sustained academic attention to the challenges faced by content moderation systems, both human and automated. This body of work has spanned across a wide variety of disciplines including, but not limited to, media studies (Carmi, 2020; Gerrard, 2020; Gillespie et al., 2020), archival research (Agostinho et al., 2019; Thylstrup, 2019), and legal studies (Cobbe, 2020; Llansó et al., 2020). While commercial content moderation systems that are developed by social media platforms often span across modalities, I specifically focus on text-based content moderation systems, i.e. NLP systems, for online abuse in this dissertation.

The academic inquiry into content moderations from the NLP community has identified several challenges with machinic content moderation systems, including poor generalisation of optimised models onto new data and contexts (Fortuna et al., 2021; Talat, 2016), socially discriminatory model predictions against already-marginalised communities (Davidson et al., 2019; Dias Oliva et al., 2021), poor ability to understand in-group communicative norms (Dias Oliva et al., 2021), and annotation biases in the data creation process (Davidson et al., 2019; Talat et al., 2018). Beyond the challenges in model performance, several key concerns have been raised about the data creation process. For instance, Wiegand et al. (2019) and Vidgen and Derczynski (2020) raise concerns about contemporary methods for collecting data, while Talat et al. (2017, 2018) point out that there is an incompatibility between widely used definitions. Vidgen and Derczynski (2020) further highlight how developing resources for English Language abuse detection has been over-emphasised by the NLP research community. Finally, applying both to modelling and data creation, the processes in the field are unable to situate speech within the context and communicative intents of the speakers (Bender et al., 2021; Dinan et al., 2020). Such a lack of ability to situate content moderation technologies within the communicative norms of *many* communities, provides privileges for those few communities that are represented when such technologies are applied across all communities. As Dias Oliva et al. (2021) show, such universality has come to mean a tacit approval of white supremacist speech while simultaneously marginalising speech from the queer community. Such consequences are unavoidable if research in content moderation technologies continues to seek universality as a solution to the issue of addressing large scale content production. For this reason, I argue that research in content moderation technologies should move from universal definitions and operationalisations of abuse, towards methods that follow the notion of “small is beautiful” (Schumacher, 1973), in which content moderation technologies centre the communicative norms of the communities that they are applied on. Although each of these challenges and directions are of equal importance and each require urgent attention, I focus my efforts first on considering the socio-political consequences of content moderation technologies, and aim to explain which modelling decisions may influence discriminatory outcomes. Second, I seek to connect the consequences of contemporaneous methods for content moderation technologies and develop new techniques for modelling abuse that aim to address some of these concerns. These twin objectives thus collectively seek to highlight the ways in which content moderation technologies are built to intrinsically optimise for universalism and white respectability politics through the political economies that they are created within (and reinforce), and the technical means of optimisation that prioritise the frequent and hegemonic over the diverse. Moreover, by jointly considering and making explicit the connection between the technical and the social of such socio-technical systems, I am afforded the ability to examine ways in which inadequacies, stemming from optimisation technologies, can be addressed from within a framework of optimisation. The over-arching research questions below then seek to first make explicit how the technical and social are connected, and rethink modes of computation to address methods that are factors in the socially discriminatory patterns of content moderation technologies.

**RQ i** *What technical and social factors are present in the socially discriminatory predictions of content moderation systems?*

**RQ ii** *In which ways can computational methods be used to address limitations that are influential in discriminatory outputs from computational modelling?*

I examine these research questions by subdividing my thesis into four sections: First, I address questions around how abuse is defined and the consequences of such definitions on the content that is subject to such systems. Second, I address model generalisability onto out-of-domain data labelled for abuse using low-dimensional data representations. Third, I examine the impacts of representing different contexts into machine learning models in terms of improvements on classification metrics. And finally, fourth, I discuss how marginalisation caused by machine learning systems is driven by modelling choices and the various ways in which designers of machine learning systems embed their own subjectivities into the models.

I choose these specific foci because my core interest is identifying how humans fit into the structure of content moderation systems, as they are currently being built. Through my interventions, I hope to identify theoretical and practical means for content moderation technologies to come to more closely respect and represent the humanity of the people who are subject to them.

To examine these questions, I begin by examining content moderation systems and logics theoretically through the field of discard studies. Specifically, I examine how people are impacted by the operationalisation of annotation guidelines and the modelling approaches chosen. Then, I consider a method for modelling abuse that seeks to address the concern of models over-fitting to highly salient terms by performing large scale vocabulary reductions. Next, I investigate how joint optimisation of abusive, and seemingly, semantically similar non-abusive tasks impact model performances for abuse detection and analyse the impacts of different classification task configurations. Finally, in a return to theory, I view machine learning and NLP through the lens of Science and Technology Studies (STS) to expose the ways in which practitioners perform a series of embodiments and disembodiments to each stage in the machine learning development pipeline. For each section, I present research questions that are subsumed by the two over-arching questions presented above. These questions allow for specific points of entry through which we are afforded the ability examine content moderation technologies as a whole. Contesting with the breadth and depth of content moderation technologies clearly requires that the research conducted has a base in a number of disciplines, from the computational methodologies to the anthropological and to the sociological, and in this case reflexive methodologies. For this reason, I address each section of my thesis from the disciplines which are best suited to answer the research questions at hand. I note here that any attempt to divorce the social from the computational, or vice versa is bound to fail to grasp the complexities and complications that they each bring, rendering the insights wanting, if not incomplete. Next, I provide a deeper description of each section of the paper, to serve as a guide to where my interests intersect with the reader's.

First, in chapter 4, I consider the nature of the task of detecting abusive content from the perspective of discard studies. Through an analysis of two content moderation systems, I examine how power differentials and respectability politics to determine the boundaries between 'the healthy' and 'the toxic' are embedded into content moderation infrastructures and expressed through them. In these analyses, I consider the aims of computational processes for abuse detection and the methods with which researchers have attempted to achieve these. Moreover, I critically examine how the notion of 'toxicity' has been operationalised in computational boundary work and the implications that these operationalisations have on the political economies of the content moderation infrastructures. Building on theories of classification and purification (Douglas, 2005) and content moderation cast a digital sanitisation practice (Lepawsky, 2019), I argue that processes and political economies of content moderation are co-constitutive of one another, thereby entrenching content moderation infrastructures within pre-existing systems of marginalisation. To address such issues, I suggest that content moderation should move beyond the question of content removal towards a productive "re-ordering of [...] environment[s]" (Douglas, 2005) to allow communities to constitute their own identities. Specifically, I argue that current practices and operationalisations of 'toxicity' for content moderation systems are rooted in patriarchal and white supremacist imaginaries of acceptability.

This chapter then seeks to provide partial answers to *RQ I* by asking:

**RQ 1** *How are notions of ‘toxicity’ operationalised and modelled, and what are their socio-political implications for content moderation systems?*

**RQ 2**

**RQ 3**

Chapter 4 is written on the basis of a collaborative research project with Nanna Bonde Thylstrup (Copenhagen Business School) and a paper has been accepted in a special issue of the journal First Monday. In this chapter, I expand on the considerations of the research project in each section and particularly further develop sections 4.2 and 4.3.

Next, in chapter 5, I turn to the technical development of new content moderation systems for abusive text. Here, I address the issue of model generalisability, over-fitting and efficiency by representing documents through the Linguistic Inquiry and Word Count (LIWC) categories invoked by the contents of the documents. LIWC is a software, and an associated dictionary, that maps a set of words to their ‘psychology relevant’ categories, that can allow its users to computationally analyse proxies to the mental states of speakers. Thus, by representing documents through the LIWC categories each word invokes, it is possible to gain some insight into the mental and emotional state of the speaker (Pennebaker et al., 2001). Consequently, models that are optimised on LIWC representations of documents are optimised on proxies for the mental and emotional states of the speaker. Moreover, the set of words that are included in the LIWC dictionary figure in much lower numbers than the raw token vocabularies or sub-words computed on the vocabulary. Thus, using LIWC can afford a low-vocabulary modelling of the emotional context a speaker is in at the time of writing.

Reducing the vocabulary size throguh LIWC has a large implication for out-of-vocabulary tokens, the likelihood of over-fitting to particular tokens, and the time required for optimising the models. Using LIWC, the vocabulary sizes are minimise by up to 99% which also introduces a question of whether this particular vocabulary reduction method is viable for the task. Moreover, the reduction in vocabulary sizes and the fact that LIWC categories act as proxies to the mental and emotional states of speakers, I experiment with the out-of-domain generalisability of models optimised on documents represented through their LIWC categories.

The research questions addressed in chapter 5 are thus:

**RQ 2** *What are the modelling implications of using LIWC to substitute the use of words and sub-words as input tokens ?*

**RQ 3**

**RQ 4**

Then, in chapter 6, I proceed further into an inquiry of the impact of context on model performance for abusive language detection tasks. Here, I use Multi-Task Learning (MTL) to jointly optimise representations of abuse classification tasks and a selection of relate tasks. Specifically, I explore new and pre-existing hypothesis on the relationships between abuse classification tasks and related tasks, and how these relationships impact modelling performance. The abuse classification tasks that I explore are hate speech detection (Talat, 2016; Talat and Hovy, 2016), offensive language detection (Davidson et al., 2017), and toxicity detection (Wulczyn et al., 2017). For the non-abusive tasks explored are sarcasm detection (Oraby et al., 2016), predicting of the basis of an argument (Oraby et al., 2015), and moral sentiment prediction (Hoover et al., 2020). Through these tasks, where each configuration sets one as the primary task and all other as auxiliary, the models are optimised to all included tasks with attention to the primary task. Selecting the tasks to include however is a challenge in itself. For this reason, I perform an ablation study investigating the influence of each task onto the primary tasks, individually and in combination with one another.

Thus, the research questions explored in this chapter are:

**RQ 3** *How do the individual and combinatory use of abuse classification and non-abusive tasks impact classification of specific forms of abuse?*

**RQ 4**

**RQ 5**

**RQ 6**

The work in chapter 6 is the result of an ongoing collaboration with Joachim Bingel (Hero I/S). All of the content produced in the chapter is new and developed specifically for the purposes of this dissertation.

In the final content chapter, chapter 7, I turn a critical lens to the proposition of objectivity in machine learning models and how this imaginary influences the machine learning-based classification of abuse. Drawing on work on subjectivity from Feminist Science and Technology Studies, I examine how human subjectivities are embodied throughout the machine learning pipeline. That is, I examine how the social and cultural meaning, that is embedded in the human experience, are also embedded in the derivatives of it, e.g. in the data that are created by humans and subsequently the machine learning pipelines that are created on the basis of these data. I argue that there is an illusion of objectivity (Haraway, 1988) through which denies the embodiments of the people whose data the models rely on, the designers of machine learning pipelines, and the potential data annotators. Consequently, such delusions of objectivity provides a barrier to developing machine learning models that are developed and deployed in an equitable manner. Moreover, I argue that this veil of objectivity is a central cause for machine learning models embodying and reproducing xenophobic white supremacist and patriarchal logics. Consequently, current machine learning models for social predictions, e.g. abusive language detection, are rooted in such discriminatory imaginaries and thus, produce discriminatory outcomes. To address this concern, I suggest that the development and deployment of machine learning models should centre the subjectivities of the people whose data make the basis of the machine learning models and the people that the models are developed for. A step to achieving this goal is to analyse how the subjective embodiments of the designers are embedded in the machine learning process, such that decisions can be made which are aligned with the subjectivities of the people using the models, i.e. machine learning models should be optimised on data from specific groups and designed for the use of those groups. The implication of these arguments is that "bias" is not a quantifiable entity that can be subject to optimisation, rather biases, or subjective embodiments, permeate the machine learning pipeline and as such the goal and work of 'debiasing' machine learning models serve to obscure the situatedness and subjective embodiments of machine learning models. For these reasons, I reflect on how my own subjectivities are embodied through the modelling choices that I make throughout this dissertation.

The research questions that I explore in this chapter are thus:

**RQ 4** *How are the subjective embodiments embedded in the machine learning pipelines?*

**RQ 5**

**RQ 6** *What are the implications of such subjective embodiments with regard to developing machine learning models?*

**RQ 7**

The work in chapter 7 extends on a collaboration with Smarika Lulz (Humboldt University), Joachim Bingel (Hero I/S), and Isabelle Augenstein (University of Copenhagen). Subsections of the chapter are currently in review at the Conference on Fairness, Accountability, and Transparancy (FAccT), 2022. The work presented here expands on the collaborative work across all sections and the sections examining the machine learning models developed in this dissertation are entirely new.

Through the work in these chapters, I examine different aspects of the machine learning pipeline for abusive language detection to gain an understanding on a) what the implications of task definitions and modelling approaches are, b) how do distinct data representations influence model generalisability and efficiency in optimisation, c) how related tasks can influence in-domain performance, and finally d) how subjective embodiments are expressed throughout the machine learning pipeline and what the implications are for developing equitable machine learning models. By examining in detail the machine learning development pipeline for abuse detection, from conceptualisation to model development, I provide detailed insights into limitations that occur at each step of the development process. The findings of my research, and the implications and considerations that derive from them, provide for a number of future directions for developing content moderation technologies. Crucially, I call for the technical research into content moderation technologies to re-orient itself around the lived realities of marginalised groups in society, i.e. those who are most at risk of harms when their content is over-moderated and content about them is inadequately moderated, such that the technologies that we develop come to serve those who are most in need of them.

Finally, before I release the reader to the core content of my thesis, I wish to provide disclaimers for the contents in this thesis. First, following Kulynych et al. (2020) I refer to *training* machine learning models as *optimising*. I make this decision because machine learning models are a specific kind of optimisation technology rather than a sentient entity which can be trained. Moreover, this distinction between *training* and *optimisation* allows us to view optimised models more clearly as statistical artefacts rather than anthropomorphised machines that “learn”. Casting aside the veneer of sentient capabilities provides space to critically examine the decisions and artefacts that influence machine learning models to produce the desired and undesired effects alike. Similarly, I refer to content moderation technologies as *machinic* (Parisi, 2009) rather than *automatic*. This distinction is made to highlight the inhuman nature of the underlying technologies, and consequently the judgements that are made by the technologies. Second, I do not make use of pre-optimised embedding layers or fine-tuned language models in my work. Although both techniques have been shown to improve performances of machinic abuse detection (Fortuna and Nunes, 2018), they are not compatible with some experiments in chapter 5 and thus influence comparability between the experiments conducted in the chapter. For the chapters in which their use would not limit direct comparability, these methods carry a range of social biases and norms that are beyond the scope of this thesis to examine. Thus I avoid using them as a way to ensure that the claims I make in this thesis are a result of the models and modelling decisions, rather than spurious correlations. Third, as I argue in chapter 7, there is a need for an active consideration of the lived subjectivities practitioners' and researchers' in computational methods. To remain consistent with my own recommendations, I use first person singular pronouns throughout this thesis. In doing so, I follow a tradition in feminist (Haraway, 1988; McIntosh, 1988; Whitson, 2017, e.g.) and Black feminist (Hill Collins, 2002; hooks, 1989; Nadar, 2014, e.g.) scholarships and critiques of the veneer of objectivity in science. Fourth, I seek to avoid providing examples of abusive texts where possible. Although my thesis is centered around content moderation and therefore requires some examples, I believe that we, as researchers, should limit the amount of abusive content in our work, as frivolous inclusion of such texts can be limiting to the potential readership and institutionalise and archive the very harms that we seek to mitigate. Lastly, beyond being a culmination of the research I have conducted, this thesis is also strongly influenced by the conversations I have had with activists and social scientific researchers, and the organisational work that I have done towards diversity and inclusion in the NLP field and organising the Workshop on Online Abuse and Harms, which I founded in 2017. Through my organisation I have come to have discussions with a variety of scholars and organisations that are addressing aspects of online abuse, including Seyi Akiwowo, Mikki Kendall, Maya Indira Ganesh, Alvin Grissom II, Kat Lo, Su Lin Blodgett, Joris van Hoboken, Seeta Peña Gangadharan, and many others. While I do not claim to speak for any of them or any particular community in this work, there is little doubt that the shape of my would be very different, and much poorer, without the discussions and conversations that we have shared.

## 1.1 Dissertation Structure

In this chapter, I have introduced the different research questions that I will be examining throughout this dissertation. Here I provide a brief overview the structure of the dissertation.

In chapter 2, I introduce some of the foundational concepts and theories that I rely on that have been developed in primarily non-computational fields.
In chapter 3, I turn to lay the computational foundation for the work conducted, introducing the various modelling aspects that I rely on.
In chapter 4, I examine how notions of 'toxic' and 'healthy' are operationalised and the implications these have on machine learning models and their outcomes.
In chapter 5, I focus on questions of model generalisability and the impacts of data represen-

tations.

In chapter 6, I examine how abuse detection tasks and other tasks related to abusive language detection can be used in a Multi-Task Learning setting to improve in-domain classification performances.

In chapter 7, I consider how machine learning models rely on the illusion of objectivity to disembody themselves, and the developers, from the subjective contexts they are embedded in.

Finally, in chapter 8, I conclude on the insights from the different chapters and suggest avenues of future research that take these insights into account.

# Chapter 2

# Theoretical Background

As abuse and automated-decision making systems for detecting abuse can disproportionately affect some populations over others. Often communities that are already disproportionately subject to negative societal biases face the worst consequences of such systems. Therefore, this dissertation considers abuse not only as a technical question of how to address abuse, but also a social question of how abuse detection systems exist and affect greater societal contexts. To fully speak to these issues, it is necessary to have a solid understanding of power and processes of marginalisation. This chapter introduces the core concepts from social science, science and technology studies (STS), and Critical Race Theory (CRT) that can provide such a foundation. Beyond this, this chapter reviews the previous literature that the latter chapters of this dissertation relies on. Speaking directly towards online abuse, I discuss work on conceptualising online abuse and hate speech. I briefly introduce the legal landscape surrounding online abuse and content moderation. Finally, as I work with data that is potentially sensitive, I provide a brief consideration of ethical concerns around using social media data for research.

## 2.1 Privilege and Marginalisation

In a consideration of how automated systems might impact different populations, it is necessary to reflect on the positions of groups in society and how they each group is enfranchised and disenfranchised, that is to consider how different groups are privileged and marginalised. The question of marginalisation and privilege has been subject to a large amount of attention from a vast number of fields, including computer science (Bender et al., 2021; Dinan et al., 2020; Mitchell et al., 2019), gender studies (McIntosh, 1988; Mohanty, 1984), law (Crenshaw, 1989), critical race theory (Benjamin, 2019; Myers, 2019), and science and technology studies (Haraway, 1988). In this chapter, and in this dissertation, I draw on the knowledge produced in critical race theory, science and technology studies, and computer science.

### 2.1.1 Theoretical Underpinnings

In 'Black Reconstruction in America' Dubois (1935) argues that whiteness in America functions as a 'psychological wage' providing poor whites in the nineteenth and early twentieth century social status due to not being Black. Through this distinction, he argues that whiteness offers a wage to poor whites who are similarly exploited by capitalism, offering 'compensation' beyond monetary compensations. Further, assigning whiteness a higher value, or an idealised position, relies on the devaluation and marginalisation of Blackness and Black existence.

Topics of marginalisation and privilege have been widely explored and is still an active area of research, for instance, contemporary scholars such as Safiya Noble (2018) and Ruha Benjamin (2019) examine how computing technologies continue to find ways of subjugating Black people to white supremacist ideals. Beyond a technological focus, these concepts have also been explored and expanded on to gender (Butler, 1990; de Beauvoir, 1953; McIntosh, 1988), sexuality (McCready, 2004), religion (Beaman, 2003), the intersection of gender and race (Crenshaw, 1989; Voigt et al., 2017), and other aspects of identity. McIntosh (1988) describes in her essay ways in which she is privileged as a white woman in comparison to Black women. Thus she highlights in very specific ways in which whiteness affords privileges and Blackness is marginalised. Throughout the many examples she list, she oscillates between highlighting larger social structures, such as *"I do not have to educate my children to be aware of systemic racism for their own daily physical protection"* and daily experiences *"If a traffic cop pulls me over or if the IRS audits my tax return, I can be sure I haven't been singled out because of my race"*. Through this description, she calls to attention how processes of marginalisation and privilege operate in both the micro and macro scale, influencing all aspects of life. Through highlighting ways in which she is privileged, McIntosh (1988) also highlights ways in which others are marginalised. It is thus clear that privilege and marginalisation can be thought of as two sides of the same coin, where one is advantaged, another may be discriminated. In more concrete and operationalisable terms, the concept of privilege describes how some demographics are not systematically and systemically disadvantaged. In comparison to marginalised communities, privileged communities receive beneficial treatment due to their distance from marginalised identities. Many social issues have been explained through the concept of privilege such as gender representations (Butler, 1990) and police treatment of Black people in the United States of America (Voigt et al., 2017).

In recent years, there has been a greater focus on social and economic disparities in context of race and gender, including not being financially punished for ones gender expression (Lombardi et al., 2002); and the increased risk of lethal encounters with law enforcement depending on racial identity (Zack, 2015).

As humans we are never confined to a single identity. In recognition of this, Crenshaw (1989) analyses three legal cases in the United States in which Black women were discriminated against while Black men and white women were not. She argues that the discrimination of Black women face marginalisation across multiple axes as they exist on the intersection of several marginalised groups: namely being Black and being women. In one of the three cases presented, Hughes et al. v. General Motors, the plaintiffs alleged that General Motors had engaged in discriminatory practices when they fired them, based on seniority. Through the trial it was discovered that all Black women hired after 1970 were fired in a round of seniority-based lay-offs. The courts found that no discrimination had occurred on the basis of sex, as white women were not fired, thus the discrimination could not be based on gender. The court proposed that the case be consolidated with another active case against General Motors on the grounds of racial discrimination. However, as the case was brought on the basis of combination of gender and race-based discrimination, the two lawsuits could not be consolidated. Thus, the marginalisation of Black women could not be identified along racialised or gendered lines individually. To expose the marginalisation faced by Black women Crenshaw (1989) theorises an *intersectional* lens. Crenshaw (1989) sheds light on the unique forms of discrimination that are rendered invisible by not considering, or refusing to consider as in the cases she discusses, how existing on the intersection of multiple marginalised identities, such as Black women, creates forms of marginalisation that are distinct from being marginalised along a single axis.

### 2.1.2 Marginalisation Through Technology

While Crenshaw's (1989) paper was published more than thirty years ago, the underlying consideration of how multiple identities intersect to create new forms of discrimination is still relevant to new technology that is being developed. In 2018, Kearns et al. (2018) independently realised that when developing algorithms to detect bias in machine learning systems it is possible to create models that are fair on the basis of singular characteristics, such as race or gender but biased towards the combination of characteristics. They thus propose an algorithm to identify an optimal number of sub-groups to consider for addressing bias in machine learning systems. In recognition of Kearns et al. (2018) and inspired by Crenshaw (1989), Foulds et al. (2019) propose a new measure of the fairness of machine learning systems that takes into account an intersectional nature of marginalisation.

However, the work of Kearns et al. (2018) and Foulds et al. (2019) both operate within the confines of the machine. As Blodgett et al. (2020) reminds us, the question of what bias and fairness mean is an inherently normative question, for which reason it is imperative that researchers define their notion of bias. Blodgett et al. (2020) recommend that considerations of bias take into account the social hierarchies that exist outside of the realm of modelling, arguing that “work analysing ‘bias’ in NLP systems will paint a much fuller picture if it engages with the relevant literature outside of NLP that explores the relationships between language and social hierarchies.” Thus, computational work cannot be divorced from the social systems that it exists in.

In another line of recent work on how technology can marginalise, Sweeney (2013) identified how search results from Google were racially biased against names that are frequently used by African Americans. Sweeney (2013) found that when searching for names with a higher association with African Americans, advertisements for examining arrest reports were shown at a higher frequency than when searching for names frequently used by white Americans. Going a step further, Noble (2018) shows how search engines reinforces racialized and gendered logics and can aid in radicalisation processes in favour of white supremacy. She argues that search engines do not merely reflect society and frequent search terms, but create their own reality through ranking.

> Search does not merely present pages but structures knowledge, and the results retrieved in a commercial search engine create their own particular material reality. Ranking is itself information that also reflects the political, social, and cultural values of the society that search engines operate within[...]
>
> Safiya Noble (2018, p.148)

Moving beyond the world of search engines, Benjamin (2019) describes how information technology at large reinforce white supremacy by providing disparate outcomes for different racial groups. Benjamin (2019) argues technology encodes such discriminatory biases in spite of, and in part due to, its ‘allure of objectivity’. She argues that within our global social structure “codes operate within powerful systems of meaning that render some things visible, others invisible, and creates a vast array of distortions and dangers” as technology operates within such systems do not require the explicit intent for racism to produce racist outcomes. In one example, she highlights the identity number system in India that produced discriminatory outcomes when their identity number, Aadhar, could not be identified:

> There are already reports of citizens being denied welfare services, including children unable to receive school lunches when their Aadhaar could not be authenticated. In this way the New Jim Code gives rise to digital untouchables.
>
> Ruha Benjamin (2019, p. 133)

Benjamin (2019) argues that this signifies a discriminatory blindness in technology that are in part brought by the makers, who are predominately white and male. Moreover, she describes a number of other examples ranging from designating Black neighbourhoods as incubators for criminality to using NLP to identify the Roman numeral ‘X’ to represent the number ten, resulting in incorrect street names such as “Malcom Ten Boulevard”. She argues that “No malice needed, no N-word required, just a lack of concern for how the past shapes the present” (Benjamin, 2019, p. 48) Notably, Benjamin (2019) proposes four dimensions that information technological systems rely on to discriminate, what she refers to as “the new Jim Code”. The first dimension is the appearance of impartiality, which she argues is not impartial given its embedding in the global social structures. The second is personalisation, which relies on the use of stereotyped information to be created. The third is merit, although systems of merit themselves are subjective and prejudiced as they too operate within our social structures. Finally, she argues that the fourth arises from “forward-looking (i.e. predictive) enterprises that promises social good” (Benjamin, 2019, p. 85), one such instance being machine learning, and as this dissertation argues, abusive language detection technologies.

## 2.2 The ‘God Trick’

> No one ever accused the God of monotheism of objectivity, only indifference
>
> Donna Haraway (1988)

In her foundational feminist STS work, Donna Haraway (1988) calls to question the notion of objectivity, critically examining science communication through a feminist lens. She argues that knowledge production is an *active* process, in which we subjectively construct knowledge based on our particular, subjective bodies. She argues that in science communication an ‘objective’ position is used to describe the object of study. However, such an ‘objective’ position, like all other positions comes with its own limitations in terms of what things are rendered visible and what is obscured. Thus, an ‘objective’ position is no less subjective, as it privileges the point of view of a particular body marked by subjective social and political meanings and possibilities along the lines of race, class, geography, gender etc. In contrast to other ‘subjective’ positions, an ‘objective’ position claims omniscience for itself by denying its own particular embodiments. Through this denial the ‘objective’ position obscures its own subjective rootedness. In the ‘objective’ position’s denial of the subjectivity of its own body, the objective position elevates itself over other ‘lesser subjective bodies’, thus playing the ‘God trick’ (Haraway, 1988). Notably, within the frame of the existence of an ‘objective’ body, as Haraway (1988) argues the ‘objective’ body is that which is held by “unmarked positions of Man and White” (Haraway, 1988, p. 8). Subsequently, the ‘lesser subjective bodies’ are those that do not fit within these, that is people of colour and women.

Through its own disembodiment, the position of objectivity claims to be ‘universal’ and free from embodied socio-political meaning and is therefore applicable in all contexts and can be imposed on all other subjective positions (Mohanty, 1984). From this it follows that embodied positions are mired in a particular, in contrast to ‘universal’, context. Accepting new knowledge from these specific embodied represents a threat to the claim of omniscience presented by the disembodied ‘objective’ position. However, as Haraway (1988) argues, subjectively embodied positions allow for things to made visible, that are otherwise rendered invisible to the ‘objective’ position. For instance, in the context of labelling uses of the *n-word*, an exclusive focus on its derogatory use would imply an understanding of the word through a disembodied and universalised position, as this universal position is often occupied by the white male body (Haraway, 1988). Only through an engagement with the particularised experiences and histories of Black bodies can the rich cultural meaning that is crafted in African-American communities be revealed and observed (Rahman, 2012).

## 2.3 Theoretical Approaches to Content Moderation

In her ground-breaking book ‘Behind the Screen: Content moderation in the shadows of social media’, Sarah T. Roberts (2019) defines Commercial Content Moderation (CCM) as professionals who are employed to “evaluate and adjudicate online content generated by users and decide if it can stay up or must be deleted” (Roberts, 2019, p. 1), in other words, to ‘clean’ internet platforms for content that is unwanted by the platforms. Roberts argues that CCMs are often outsourced to call centres in the global south and boutique firms in North America, with a minority work force held in-house by social media companies (Roberts, 2019). For outsourced workers, the work is frequently poorly paid and has steep psychological costs to the workers undertaking the job of keeping social media companies palatable to their core audiences. As one interviewee states “Horror movies are ineffective at this point. I have seen all that stuff in real life” (Roberts, 2019, p. 122). Moreover, she argues that large companies reproduce colonial logics by creating special ‘ecozones’ in Manila, developed in part to respond to the needs for “uninterrupted electricity, the capacity for large scale bandwidth for data transfer, and so on" (Roberts, 2019, p. 183). In these zones, call centre workers are increasingly given higher targets and smaller workforces (Roberts, 2019, p. 178) due to the increased competition from other companies in other parts of Asia. Thus, large social media companies establish their own colonies of exploitation in such ecozones where workers' rights are competing with the risk of companies outsourcing to a different company in a different part of Asia.

In her work on theories of social pollution, Mary Douglas (2005) examines how meaning and community are made through the positive reordering of the environment to separate the subjects and objects that belong and those that do not, i.e. what is not and what is dirt. She argues that "dirt is the by-product of a systematic ordering and classification of matter, in so far as ordering involves rejecting inappropriate elements" (Douglas, 2005). Dirt is then not an independent attribute of an object or subject, but a "residual category rejected from our normal scheme of classifications" (Douglas, 2005). This classification between what is dirt and what is not, practised through rituals and habits that create coherence within communities, helps establish borders between what and who belongs in a group and what does not. As communities reject the impure, members of the collective, or in fact society, form a shared meaning. Through these processes of meaning-making the collective can maintain its integrity. Dirt then becomes something communities avoid in order to prevent the breakdown of meanings. Thus to Douglas (2005), the avoidance and removal of dirt are not negative processes of removal, they are a "positive effort to organize the environment" (Douglas, 2005) of the community in which the removals takes place. Further she argues that we seek to re-order our environment to make it conform to an idea.

As the identification, demarcation, and expulsion of dirt are collective actions, the definitions and understandings of what constitutes dirt is subjective in nature and meaning can only be attributed within a given system: "no single item is dirty apart from a particular system of classification in which it does not fit" (Douglas, 2005). Further, in her conceptualisation dirt is an encompassing label for "all events which blur, smudge, contradict or otherwise confuse accepted definitions" (Douglas, 2005). Dirt is then contextual, and what is dirty in one situation may not be dirty in another. She exemplifies this through the mundane: food is not necessarily dirty, "but it is dirty to leave cooking utensils in the bedroom" (Douglas, 2005).

Given the contextual nature of dirt, many might imagine that they are unequivocally able to identify dirt, Douglas (2005) argues that detecting dirt is complex as with the contextual nature of dirt, it follows that there can be no such thing as absolute dirt or clean, as these depend on the subject observing it. Thus, what is dirt to one might be valuable to another.

Conversely, as Lepawsky (2019) argues, the positionality of discarding, or cleaning, may result in discarding or maintaining things that are valuable to one, but are harmful to another. Who then gets to exert such power to determine what stays and what remains becomes a question of differential power relations. Indeed, Hall (1997a) similarly ties the classification of dirt and the clean to racist logics of social purification:

> What you do with dirt in the bedroom is you cleanse it, you sweep it out, you restore the order, you police the boundaries, you know the hard and fixed boundaries between what belongs and what doesn't. Inside/outside. Cultured/uncivilised. Barbarous and cultivated, and so on.
>
> Stuart Hall (1997a, p.3)

By investigating the question of power relationships and commercial content moderation Lepawsky (2019) extends Douglas (2005) framework into digital spaces, arguing that such work can help us understand online communities as systems that must constitute themselves through removal, for which human and automated content moderation systems act as the filters that allow for such constitution.

In Hall's (1997a), theory of encoding and decoding, it is argued that expressions are written, or encoded, with a specific understanding which may differ when it is interpreted by the reader from one of three positions: dominant, negotiated, or oppositional. These moments of interpretation create a space for uncertainty and instability. Where things have been encoded with one meaning, they can be decoded with a different meaning, e.g. a semi-colon followed by a close parenthesis may be encoded as an indicator of sarcasm, but decoded as ungrammatical. Such oppositional reading can give space for subcultural communities, that stabilise their own meaning-making process such that the community understands a semi-colon followed by a close parenthesis as indicating happiness.

In considering Hall (1997a) and Lepawsky (2019), the logics of dirt cannot be disconnected from the oppression experienced by marginalised people. As Risam (2015) notes, toxic "has become cultural code for irritants and pollutants that disrupt our lived experience". Risam (2015) argues that discourses of toxicity are invoked in cultural conversation between hegemonic and marginalised bodies, and are weaponised against marginalised bodies through an engagement in 'toxic slippage'. Toxic slippage denotes when, in response to 'toxic' behaviour, users, or this dissertation theorises, computational methods, engage in toxic behavioural patterns. Camp theory, as Schaffer (2015) argues, can offer a "useful mode of reading for *any* field of study marked by questionable binaries". As the question of 'toxic'/'non-toxic', 'abuse'/'non-abuse', and 'hate'/'not-hate' provide such questionable binaries between the desired and undesirable, camp theory's focus on re-evaluating a culture's 'trash' offers a well-established theoretical frame to re-articulating such concepts, complicating the ways in which content "can blur and transgress and cover in glitter those boundaries between waste and not-waste [...] without pretending that waste has stopped being waste"(Schaffer, 2015). As camp provides a mode of queering the understanding of waste from the discarded to the celebrated, it can be read as a "queer way of knowing, one that emphasises reader relations over any inherent meaning of a cultural object" (Schaffer, 2015).

A camp reading of content moderation systems and the decisions produced by them, would centralise the experiences of the marginalised individual faced with consequences of toxic slippage and how dirt must apply to them. Such a reading would fortify the centrality of the subjective positionality of the individual in relation to the subjective positionalities of the content moderation system.

## 2.4 The Legality of Abuse

As the detection of abuse may be closely linked to the content moderation and the removal of content, any consideration of tools with the purpose of identifying abuse must also consider the legal context context in which they may operate, to situate the work in computational content moderation within the social and political realities that it exists in. Here, I consider two forms of legal contexts: 1) platform policies and 2) regulatory frameworks and their influences on using automated methods for detecting abusive language.

### 2.4.1 Moderation Practices

To understand how content moderation can work, it is necessary to consider the communicative power of frameworks for reporting inappropriate content. Crawford and Gillespie (2016) argue that reporting mechanisms constitute a communicative channel between users and the platforms. Thus, the question of content reporting becomes a more nuanced space than the question of what content is permissible, instead it becomes a question of what is desired by different user communities (Crawford and Gillespie, 2016). Simultaneously, reporting frameworks also allow for companies to control the expressiveness of the communication between users and the company and the volume of this communication (Crawford and Gillespie, 2016). On one hand the expressiveness of the reporting can be controlled by the level of detail users are afforded when reporting, on the other hand, volume can be controlled through the ease of access to the flagging mechanism. When platforms set the degree of expressiveness of the flagging mechanism, the detail allowed provides a signal on whether the platform are interested in the particular way content offends or simply that users find that the content should be disallowed (Crawford and Gillespie, 2016).

While often a flagging is often a solitary effort performed by a single user, it can also be the result of a strategic means of communication by a coordinated user group. For instance, a group of bloggers angered by pro-Muslim content on YouTube started the 'Operation Smackdown' campaign in 2007 to remove such content. In this coordinated attack, coordinating users created lists of YouTube videos for other users to flag as 'promotes terrorism' (Crawford and Gillespie, 2016). Here flagging campaigns are used to reinforce social hegemonies.

On the other hand, when platforms remove content that is socially acceptable to a large majority, there can be strong backlashes against the platforms. In one such event, Facebook removed an image of two male actors kissing on the television show EastEnders. Following this removal there was a large outrage, accusing Facebook for perpetuating homophobia as images of straight couples had not been subject to removal. Following this controversy, Facebook reversed their decision to remove the image and apologised for the removal (Crawford and Gillespie, 2016).

### 2.4.2 Platform Policies

Many large social media companies lay out their policies for acceptable behaviour on their platforms, often detailed in their user guidelines. Many of these have similar phrasing on the acceptance of abusive language,[1] in figs. 2.1 to 2.3, I show excerpts of the policies on hate speech and prohibited content of three social media platforms.

These excerpts highlight how platforms envision what prohibited content appears on their platforms and which priorities that they have. For instance, in Facebook's policy (see fig. 2.1) they make a brief mention of real-world harms, however incitement to violence is omitted from their guidelines. Twitter on the other hand (see fig. 2.2) provide a much less clear in terms of what is prohibited and what is allowed through the ambiguity that they use (Kirtz et al., 2022). It is, for example unclear what constitutes are 'direct attack' on diffuse concepts such as demographic groups. Lastly, Reddit (see fig. 2.3) very succinctly describe the kind of content that is prohibited on the platform, so succinct that considerations around hate speech are entirely omitted.

[1] For the full policies see `https://en-gb.facebook.com/communitystandards/hate_speech` for Facebook's policy on hate speech, `https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy` for Twitter's policy on hate speech, and `https://www.redditinc.com/policies/content-policy` for Reddit's policy on prohibited content.

In all three excerpts of the policies, there is a prohibition of content which attacks others, in Reddit's policy on encouraging and inciting violence they further outline that users should not post content that "that encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group of people"[2] establishing similar outlines for acceptable conduct as seen on Twitter and Facebook. Facebook note in their Community Standards Enforcement Report[3] that they acted on 4 million items for hate speech and 2.8 million items for bullying and harassment in the period of January to March 2019. The report does not detail the number of user reports received, nor the amount of content which was not removed.[4] Considering the scale of the items which have actions taken for different kinds of abuse, there is an incentive to allow for some automation to guide the attentions of human moderators or decrease the volume of content that moderators need to consider.[5]

While Facebook report high numbers of removals, the performance of their (human and automated) moderation practices have been the source of criticism as a number of activists have reported being temporarily banned for speaking about racial discrimination while abuse and discrimination received is not addressed (Sharif, 2019). Moreover, users have been noting that simply talking about race may mean that their post is removed, particularly if the poster is not white (Guynn, 2019). A report from Pro Publica details that the policies of Facebook in determining whether a post violates community guidelines, by seeking global standards, effectively "protect the people who least need it and take it away from those who really need it." (Angwin and Grassegger, 2017)

Perhaps most damningly, public figures are often exempt from community guidelines with the argument that their content is in the public interest (Díaz and Hecht, 2021), in spite of the fact that public figures have the potential to influence much larger groups of people, and incitements to violence, for instance, can reach a much larger group of people.

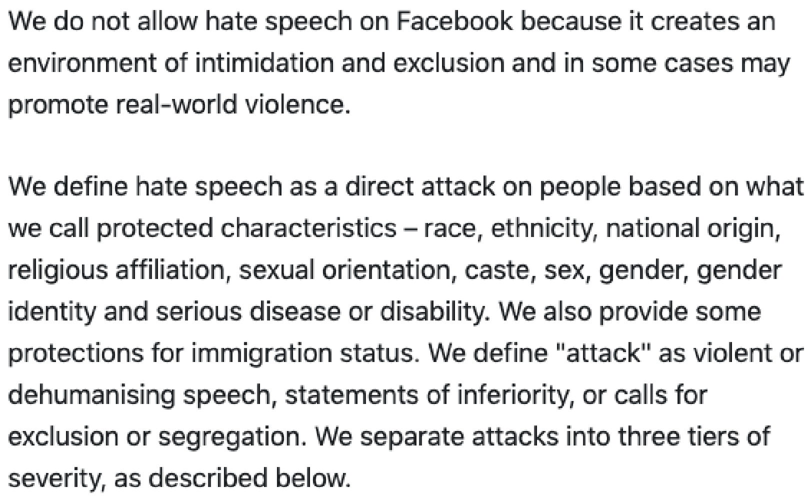
Considering how such policies are described, understood, and enacted by commercial content moderators, content moderation algorithms, and users, Kirtz et al. (2022) perform an analysis of how different platforms describe their community guidelines and their downstream implications for content moderation technologies and user experiences. They argue that the platform policies are often ambiguous and do not provide a clear expectation to users as

[2] for the full policy see `https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-post-violent-content`

[3] See report here: `https://transparency.facebook.com/community-standards-enforcement#hate-speech`

[4] Acted on here means acknowledging that the content does violate community standards and an action was taken by Facebook.

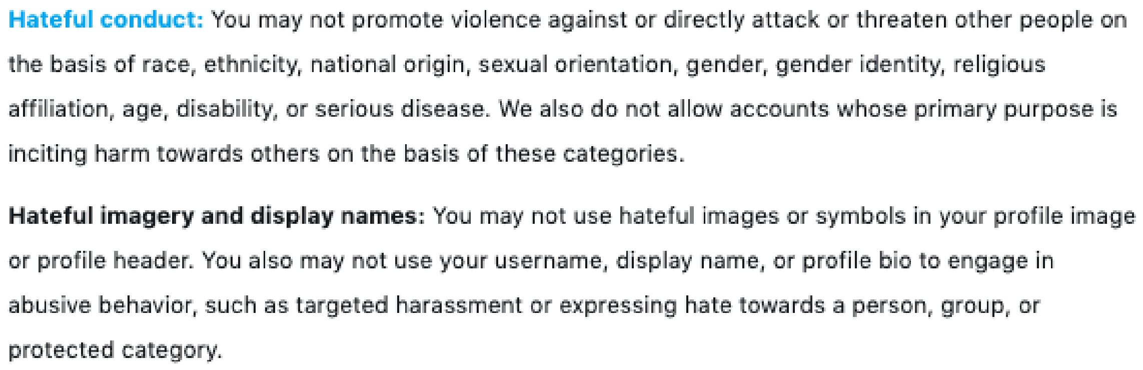
[5] The Community Standards Enforcement Report details that automated systems are deployed but do not detail the performances of the system in terms of *accuracy, precision*, or *recall*.



We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence.

We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation. We separate attacks into three tiers of severity, as described below.

**Fig. 2.1** Excerpt of Facebook policy on prohibited content and hate speech.



**Hateful conduct:** You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

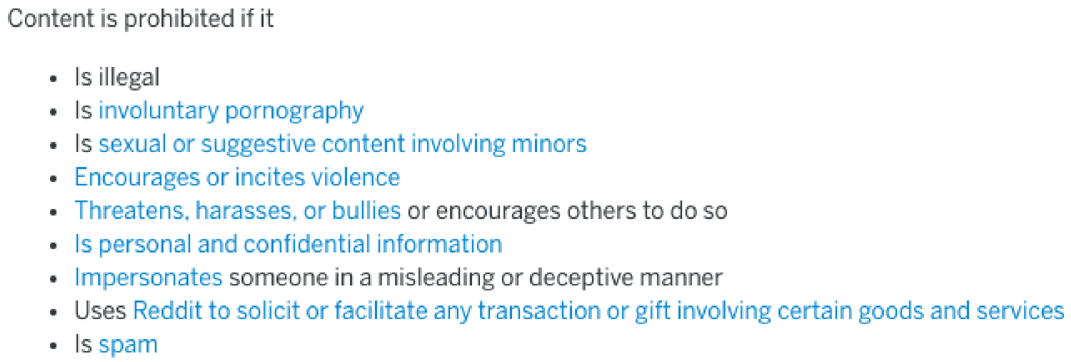
**Hateful imagery and display names:** You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.

**Fig. 2.2** Excerpt of Twitter policy on prohibited content and hate speech.

to what is prohibited content and what is allowed. Kirtz et al. (2022) further argue that the machine learning algorithms that underpin many automated content moderation systems require a degree of specificity to function that is not apparent from the policy guidelines or otherwise communicated. Developing guidelines for machine learning systems on the basis of such policies, as a number of studies in NLP do (Davidson et al., 2017; Qian et al., 2019, e.g.) then provides a large ambiguous space for human workers which makes understandings of each category tacit rather than explicit, regardless of whether the workers are employed on low wage contracts in South East Asia (Roberts, 2019) or hired by academic researchers.

### 2.4.3 Regulation

In recent years several different governments have sought distinct methods to address the issue of online abuse. For instance, the British Home Office (Home Office, 2016) and the



Content is prohibited if it

- Is illegal
- Is involuntary pornography
- Is sexual or suggestive content involving minors
- Encourages or incites violence
- Threatens, harasses, or bullies or encourages others to do so
- Is personal and confidential information
- Impersonates someone in a misleading or deceptive manner
- Uses Reddit to solicit or facilitate any transaction or gift involving certain goods and services
- Is spam

**Fig. 2.3** Excerpts of Reddit policy on prohibited content and hate speech.

European Commission (European Commission, 2016) have provided guidelines and calls to action for social media networks, while the German government passed the Network Enforcement Act (NetzDG) (The Bundestag, 2017) which aims to provide direct regulation for the moderation of online abuse and misinformation. NetzDG incentivises the moderation of content through fines if social media companies *systematically* fail to remove within 24 hours.

Building on this, the European Union is currently considering regulation on disseminating terrorist content which may have implications on hate speech and how social media platforms deal with issues such as hate speech (European Commission, 2018). The proposal contains a requirement for social media networks to remove content within 1 hour of receiving notice from a *trusted authority*.[6] In response to this proposed regulation, the European Union Agency for Fundamental Rights (FRA) highlight the difficulty in the border work in distinguishing between content that promotes terrorism, documents war crimes, or is simply hateful. The FRA suggest that the proposed regulation should provide a clear definition of terrorist content, which is limited to inciting or promoting terrorist activities, or providing instructions for making or using weapons (European Union Agency for Fundamental Rights, 2019).

In contrast to developments in Europe, the governance surrounding the moderation of online content in the United States of America is centred around section 230 of the Communications Decency Act of 1996 (Communications Decency Act of 1996, 1996). Relating to the moderation of online content, Section 230(c) of the Communications Decency Act specifies that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider"

[6] Trusted authority was not specified at the time of writing.

(Communications Decency Act of 1996, 1996), thus specifying that service providers, e.g. social media networks are not liable for the content that is created by others and hosted on their platform. The over-arching goal of Section 230 has been to ensure that service providers are not liable for hosting content that is beyond their control. However, sub-paragraph 2(a) notes that

> No provider or user of an interactive computer service shall be held liable on account of—any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected.

Thus allowing social media networks to remove and moderate content from their platforms. Section 230 has been central to the evolution of social media networks and their content moderation infrastructures, as many such networks are founded and legally based in the United States of America and therefore subject to its laws.[7] [8] Section 230, its consequences, and potentials are the deeply complex. Providing a detailed overview of the issue is therefore beyond the scope of this thesis. Interested parties can see research and organisations that more specifically deal with the legal landscape for online intermediaries Appelman et al. (2021); Díaz and Hecht (2021); Llansó et al. (2020, e.g.) and section 230 specifically Ardia (2010); Defterderian (2009); Foundation (n.d.); Goldman (2018); Leary (2018); Sheridan (1997, e.g.).

Contrasting the permissive nature of Section 230, the United States of America recently instituted two bills, the Stop Enabling Sex Traffickers Act (SESTA) and Fight Online Sex Trafficking Act (FOSTA) which take more active approaches to ensuring that online service providers remove content that offends the bills. These two bills aim to prevent sex trafficking by removing online content for sexual solicitation. As a consequence of these bills, a number of online platforms entirely removed all sexual content which particularly has adverse effects for already marginalised communities. These bills have been heavily criticised for flaws in their conceptualisation (Romano, 2018), their impact on criminal investigative work (Q, 2018), and their consequences for sex workers, which include missing persons and deaths of

[7] This seeming tension between not being liable while being given authority to exercise content moderation is also at odds with the highly permissible notion of Freedom of Speech that is institutionalised in the United States of America's law. In the law of the United States of America, the concept of freedom of speech has very few limits, in contrast to the European notion of freedom of speech, which has much greater restrictions (Banks, 2010).

[8] I note here that although the Commnications Decency Act only applies to the legal territories of the United States of America, it is often used as the guiding principle from which content moderation decisions are motivated.

sex workers (Blue, 2018; Simon, 2018). More widely, there are three main implications of the bills:

1. The bills seek to undermine (Romano, 2018; Stryker, 2018) Section 230 of the Communications Decency Act (CDA) by holding that computer service providers are not publishers and thus not liable for content on the platform (Foundation, n.d.),
2. the bills are effective retroactively (Stryker, 2018), therefore necessitating moderation of both new and historic content, and
3. the bills do not require a request to remove content prior to potential consequences for not upholding the laws.

While enforcing more regulation and shifting liability for content on platforms onto the online social networks is in line with the European and German regulations, the consequences of SESTA and FOSTA go a step further, as the bills are retroactively effective, for which reason online online platforms sought to remove all sexual content to avoid the risk of fines at the cost of people's freedom of speech. In contrast, the German and European efforts largely allow for content to remain, as platforms only need to react if and when content is reported - moreover, they are not retroactively effective, meaning that platforms did not need to perform sweeping removals. The contrasts between the European approach and the approach taken by the United States of America, have strong implications for how online platforms develop technology for content moderation. For instance, the European approach incentivises social media networks to implement robust content moderation infrastructures, that somewhat evenly balance *precision* and *recall*. The approach taken by the United States of America on the other hand, incentivises a much less precise method, as platforms have greater risk from retaining content than simply remove it all. As a consequence, several platforms removed entire sections of content created by users that *may* have had a *perceived* risk to the platform. One such platform was Tumblr which removed a large section of historical and new content produced by queer communities, due to the risk that any content related to sexuality may also have been solicitation. The work in this thesis thus more closely aligns with the European approach, seeking not to remove content from entire communities but rather identify specific instances of content to be removed.

## 2.5 Ethical Considerations

When dealing with a subject like online abuse and content moderation, there are a number of ethical issues that are necessary to bear in mind, from the anonymisation of already-existing data to the methods of collecting new data. These issues are further complicated by the strength with which identity intersects with whether something is offensive, as some things are only visible as offensive, or non-offensive, when one has intimate knowledge and familiarity with the target. What is the ground truth then becomes subject to the particular subjectivities of an individual who is involved in the communicative act.

Considering first the issue of anonymisation, though it is important to perform some level of anonymisation for ethical as well as technical reasons, i.e. minimising tokens to overfit to, there is also a greater risk of performing anonymisation to a strong degree, i.e. make the author of a text entirely unrecoverable. One such risk of strong anonymisation, which would require that the text is reordered, is that the dialectal indicators of belonging to a demographic would be erased, thus directly contradicting the need to have intimate grounding in the subjectivities of a given speaker. Moreover, when releasing a new dataset, researchers need to address the issue of potential harms arising from data being publicly available. One alternative to such publication of data is for data of a sensitive nature, i.e. labelled or identifiable social media data, to never be released publicly. An imaginable configuration for this could be that data is never released, however researchers can submit models and code to optimise and evaluate on the dataset. Such a governance model however comes with several drawbacks: 1) it is not possible for researchers to analyse the kinds of mistakes that machine learning models make, 2) the governance model severely inhibits who has access to data, and 3) it renders invisible analysis and drawbacks of the datasets, something which is of vital importance as our understanding of abuse and hate speech continuously grows.

Developing new resources for abuse detection also provide concerns and challenges. Creating a dataset for online abuse requires designating *someone* to look at a large volume of data, in which abusive content will likely appear without warning. Being situated in such a context can have mental health concerns due to the exposure to vicarious trauma, i.e. trauma occurring from witnessing traumatic events. As Roberts (2019) notes, the human workers who deal with moderating content for large tech companies often experience severe mental health issues as a consequence to their job. This issue of subjecting workers to trauma is then further exacerbated when we consider the interaction between correctly identifying online abuse and identity. Specifically, people will be most attuned to the various ways in which hateful and offensive content is directed towards their own identities. Consider for a moment dog-whistles (Drakulich et al., 2020), the coded language that is used to target specific demographic groups, the specific group that is being targeted may be the only group that can correctly identify a speech act as abusive or hateful. Thus, it is of great importance that the workers that are recruited label data about their own identities, however this too comes with a great personal cost. Recruiting people to annotate abuse and hateful speech targeted towards their own identities also means exposing them to violence that specifically targets them (Boeckmann and Liew, 2002), which may make the issue of vicarious trauma even more acute.

Turning lastly to the use of modelling and predicting online abuse, there are risks of harms to people from model misclassification, in particular to people who are in communities that are already marginalised. As Talat et al. (2018) indicate and Davidson et al. (2019) and Dias Oliva et al. (2021) lay clear, there are significant racialised issues with classification of abuse, leading to African American English speakers and queer folks being censored through a disproportionate amount of false positives. One might turn to the question of whether computational modelling is at all appropriate for content moderation or whether it is best to return to entirely human content moderation pipelines. This question is complicated by the sheer scale of content being generated, with billions of new tweets, Facebook and Reddit posts, and YouTube videos and comments created each day. The scale is then far beyond what could be hoped to be reviewed by human content moderators alone - suggesting a need for some form of automation to aid the human workers in their task. But is it acceptable for anyone to have to be subjected to the violence of harmful and abusive content at scale? If it isn't acceptable, then it is imperative that we develop content moderation technologies quickly and robustly, to address the right to be free from persecution in online spaces.

### 2.5.1 Ethics Statement

As this dissertation deals with data that is published by individuals who very often are private citizen, it is necessary to consider different aspects of the ethical use of social media data use. As I only make use of previously published data, informed consent cannot be obtained directly, as I do not have access to the necessary contact details. Moreover, several of datasets do not provide the user information to even consider access. Finally, as many of the datasets are several years old any contact information that could be gleaned from the data sets have, in many cases, decayed.

As informed consent is not sought for the data, the need for a consideration of anonymity and privacy is only heightened, to ensure that private citizens are not exposed additional harms. For this reason, all experimental parts of this dissertation have undergone ethics review for risks and harms and been approved for study. All experimental work presented exclusively makes use of previously published, and currently public datasets. Some data in these datasets are provided entirely anonymised while others are entirely de-anonymised.

In all experimental work, I anonymise all data where appropriate to avoid the undue risk of harms to data subjects.

## 2.6 Summary

In this chapter, I have introduced several key notions and concepts that will lay a theoretical and philosophical foundation of my work. Specifically, this chapter introduces the notions of privilege and marginalisation and previous work on how these concepts relate to computational techniques. I further provide a background to theory surrounding content moderation. Finally, I provide a consideration of ethical use of social media for research and conclude the chapter with a brief overview of two legal aspects of abuse.

# Chapter 3

# Computational Background

This chapter introduces related work in Natural Language Processing (NLP) and theoretical background on the machine learning methods that I use throughout this dissertation.

## 3.1 Abusive Language Detection

In recent years, the computational study of online abuse has seen a rapid increase in the number of papers dedicated starting with a handful of papers prior to 2016 to a thriving research field with numerous papers, shared tasks, and workshops (Vidgen et al., 2020a). In spite of the growth in research dedicated to the detection of online abuse, the research field is still in its infancy with a number of open questions, including questions around definition of the task, annotation guidelines, and modelling techniques. Due to the relative infancy of the field, the 'early' and 'recent' work overlap each other in time. I use 'early' and 'recent' work to distinguish 'early' from 'recent' by conceptualising 'early' as work which laid the foundations of the field and 'recent' as work that develops, or critiques, these foundations. The earliest work in the field sought to address questions of cyber-bullying (Chen et al., 2012; Daegon Cho, 2013; Reynolds et al., 2011) and profanity (Sood et al., 2012a,b) with sparing focus on demographically specified abuse, such as racism, sexism, and anti-Semitism (Warner and Hirschberg, 2012). More recently, work on demographically specified abuse has surfaced as an independent task (Gorrell et al., 2018; Karan and Šnajder, 2018; Meyer and Gambäck, 2019; Palmer et al., 2020; Park and Fung, 2017; Safi Samghabadi et al., 2017; Stoop et al., 2019; Talat, 2016; Talat and Hovy, 2016; Tulkens et al., 2015; Vidgen et al., 2020a). As a consequence of increased visibility of hate speech and abuse on online platforms, the academic inquiry into the computational detection has grown along with the regulatory responses (European Commission, 2016; The Bundestag, 2017).

Early, and contemporary computational work, has seen a great deal of focus to central questions around the task: how do we annotate and create datasets (Talat and Hovy, 2016; Talat et al., 2017; Vidgen et al., 2020a) and understanding annotator interaction and performance (Ross et al., 2016; Talat, 2016; Vidgen et al., 2020b). Early work focused on questions of marginalisation and oppression, for instance through the work of Talat and Hovy (2016) who base their annotation on works in gender studies and critical race theory, and collect data based on gendered and racialised abuse; more recently data collection and annotation processes have moved towards a demographically blind process. Such early work was inspired by the marginalisation of certain bodies and the desire to develop computational tools to protect marginalised people (Warner and Hirschberg, 2012). More recent work has instead directed its focus to demographically blind approaches to data collection and annotation, succumbing to 'marginalisation-blind' annotation processes and guidelines. Although processes that do not take marginalisation into account, but instead seek to treat every group equally provide an allure of fairness, they also encode dominant discourse on abuse with the subsequent result of the resistance to oppression and marginalisation being treated the same as marginalisation. In concert with the growing evidence of racially biased content moderation tools (Davidson et al., 2019; Talat et al., 2018), demographically blind annotation criteria and data curation pose a threat to the goal of developing tools that aid in ensuring people from the right from persecution. One such example is presented by Salminen et al. (2018) who develop a taxonomy that includes 'anti-white' as a target of hate on par with anti-Black hate in spite of whiteness as a hegemonic entity that marginalises (McIntosh, 1988). A result of this are egregious annotation choice, such as "The white will always steal; FUCK YOU TO ALL WHITES RACIST" labelled as hate speech (Salminen et al., 2018), in spite of the comment speaking to ongoing racism and the historical exploitation enacted by white societies (e.g. the theft of cultural artefacts from colonised territories (Frost, 2019), the numerous genocides committed by imperialistic colonial states (Weisbord, 2003), and the theft of bodies in the transatlantic slave trade). Moreover, and perhaps of even greater concern, the annotation and curation processes of Salminen et al. (2018) result in data responding to the abuse of authority committed by police as hate, in one such example they identify the following comment as hate "did to that poor guy. 10 s of pepper spray directly into the face, run over foot etc. equal it up a little bit, except for the detail of having a fucking stroke. So it still wouldn't be exactly what the guy went through. Fucking discusting. They get a hard on power tripping others. They are just fucking cowards", in all likelihood due to the aggressive nature of the comment.

Through such demographically uninformed processes of curating and making data, a danger of erasure of past and ongoing marginalisation and responses to it as well as critical responses to the violent abuse of authority as ‘hate speech’ that should be subject to content moderation. The question of automated hate speech detection thus, for works such as Salminen et al. (2018) is no longer ensuring the right to not be persecuted but instead insuring that processes of marginalisation remain unchallenged. For these reasons, I use the datasets released in early work, specifically I use the *Offence* dataset (Davidson et al., 2017), *Toxicity* dataset (Wulczyn et al., 2017), and *Hate Speech* dataset (Talat and Hovy, 2016) in all computational chapters. For chapter 5 which examines the influence and generalisability of vocabulary manipulation, I also use the *Hate Expert* (Talat, 2016) and the *StormFront* (Garcia et al., 2019) datasets. Each of these datasets share the common attributes that they are collected either from spaces that are hateful towards marginalised groups or have considerations of marginalisation encoded into the annotation guidelines. In chapter 6 I also use three datasets that are labelled for abuse but instead to tasks that are seemingly related. First, I use the *Argument Base* dataset (Oraby et al., 2015), the second dataset (*Sarcasm*) is developed for sarcasm detection (Oraby et al., 2016), and the final dataset, *Moral Sentiment*, examines the moral sentiments expressed in tweets (Hoover et al., 2019).

In an early effort to address issues of annotator biases and under-sampling of some forms of data in the data curation process, Talat et al. (2017) propose a typology of abuse that aims to categorise abuse by how it is characterised rather than determining the exact form of abuse. To this end, Talat et al. (2017) present a 2-dimensional typology of hate; the first dimension operates along implicit and explicitly expressed hate. Implicitly communicated hate, Talat et al. (2017) argue is hate that is communicated through subversive means by using code words and communicating implicit biases. Explicit abuse on the other hand is explicit in its intention to abuse, e.g. through the use of slurs. The second dimension concerns itself with the target of abuse that can either be a generalized other, or a specific group, the former category detailing abuse that is targeted towards small groups and individuals while the latter is aimed at generalised targets, e.g. larger demographics. It’s important to note that content may be simultaneously explicit and implicit, directed and generalised (Talat et al., 2017). For instance, content that implicitly targets Muslims, may simultaneously explicitly target a specific group of women.

Modelling for automated hate speech detection has also undergone a development from early to contemporary work. Early work was primarily focused on feature-based modelling (e.g. Davidson et al., 2017; Sahlgren et al., 2018; Talat, 2016; Talat and Hovy, 2016) whereas subsequent work has directed a greater attention to neural network based approaches (e.g.

| | *Explicit* | *Implicit* |
|---|---|---|
| *Directed* | “Go kill yourself”, “You’re a sad little f*ck” (Van Hee et al., 2015),<br>“@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga” (Davidson et al., 2017),<br>“Youre one of the ugliest b*tches Ive ever fucking seen” (Kontostathis et al., 2013). | “Hey Brendan, you look gorgeous today. What beauty salon did you visit?” (Dinakar et al., 2012),<br>“(((@User))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles” (Hine et al., 2017),<br>“you’re intelligence is so breathtaking!!!!!!” (Dinakar et al., 2011) |
| *Generalized* | “I am surprised they reported on this crap who cares about another dead n*gger?”, “300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!” (Nobata et al., 2016),<br>“So an 11 year old n*gger girl killed herself over my tweets? ^̲ ˆ thats another n*gger off the streets!!” (Kwok and Wang, 2013). | “Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home.” (Burnap and Williams, 2015),<br>“Gas the skypes” (Magu et al., 2017),<br>“most of them come north and are good at just mowing lawns” (Dinakar et al., 2011) |

**Table 3.1** Typology of abusive language presented by (Talat et al., 2017).

Badjatiya et al., 2017; Gambäck and Sikdar, 2017; Kolhatkar et al., 2020; Talat et al., 2018). In this dissertation I follow a similar pattern of developing baseline models from feature-based models and suggest neural network architectures as extensions and improvements on these. In early work, Logistic Regression (LR) and Support Vector Machines (SVM) were the most frequently used models. As the scholarship has developed, specific types of neural networks have come to dominate the modelling, namely Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM) networks. In each chapter, I perform the review of the models that are pertinent to the work in the chapter. Here instead I provide a theoretical overview of the models, their components (e.g. dropout and activation functions) and their intended functionalities (i.e. the kind of data that they are designed to operate on and which assumptions are built into the model architectures).

### 3.1.1 A Word on Definitions of Hate

The definitions of hate speech and offensive content that are used in the machine learning literature deviate quite strongly from how these concepts are defined in fields such as critical race theory, gender studies, and feminist science and technology studies. Here, I want to make clear the distinctions between these to schools of thought. The primary difference in the critical and computer scientist reading of hate speech and abuse is the consideration of social circumstance and systems of power. Critical race theory, gender studies, and feminist science and technology studies all consider how social constructs, such as race and gender influence our experience of the world. Turning this consideration of social constructs and power structures towards the question of racism, sexism, and hate speech in general, these fields then distinguish how targets are situated in terms of power when considering whether something is hate speech or not (Ging and Siapera, 2018). For instance, should two people with different identities be targeted with the same abusive statement, the impacts of that abuse will differ and speak into different histories. Should one of these people have a hegemonic identity while the other belongs to an oppressed group, e.g. one is a man and the other a woman, then the abuse will speak into histories of power and marginalisations, respectively. Following this line of thinking, abuse then is also a matter of structural power. The distinctions between discrimination and hate speech, for instance, is then co-constituted by what the specific dimension of the abuse is, e.g. gender or race, and the whether the recipient is marginalised along that axis. Thus, racism moves from discriminatory statements into hate speech, when the target is also marginalised along the axes of race, and social class-based abuse goes from discriminatory statements into when the target is also oppressed along the axis of class. It is important to note here that the distinguishing factor is not whether there is an impact or the strength of the impact of a discriminatory moment on an individual, but rather the historical and contemporary contexts within which such a moment exists.

The computer scientist and machine learning literature however largely disengages with questions of power structures. Instead, the literature tends to focus in on micro-instances of conflict and to a large degree disconnects the identities of the target with the abuse. For instance, in the conceptualisation of Davidson et al. (2017) and Wulczyn et al. (2017) the *offensive*, *hate*, and *toxic* are entirely focused on the conversational moment and the presumed intent of the speakers.[1] This approach also is divorced from the systemic reading of hate speech which makes the impact central, rather than the intent of the speaker (Ging and Siapera, 2018). The computer scientific approach to hate, is a comparatively simplistic approach which collapses important distinctions such as how people live at the intersection of multiple identities, i.e. queer brown men or working-class white women.

In the following chapters, I consider the readings which are most appropriate for the content of the chapters. In chapter 4, I consider consequences of the simplistic view computer science takes on hate speech. In the next two chapters, chapter 5 and chapter 6, I focus on the computer scientific notions of the superset *abuse* and its subsets: *Hate speech, toxicity, and offense*. Finally, in chapter 7 I return to the notion of hate speech in the systemic sense, as I consider the wider implications of the processes through which we develop machine learning technologies.

[1] Presumed by annotators.

### 3.1.2 Datasets

Here I provide an overview of the different datasets that are used throughout this dissertation.[2] For each dataset, I introduce the curation rationale, the source of the datasets, the annotation guidelines, annotator selection, and finally how each of these dimensions influence the resulting dataset. In subsubsection 3.1.2.1 I describe the datasets annotated for hate speech and abuse and then in subsubsection 3.1.2.2, I turn to the datasets used for the auxiliary tasks for chapter 6.

#### 3.1.2.1 Hate Speech and Abuse Datasets

To perform any machine learning modelling of abusive language it is necessary to use some datasets. However, as we saw in section 3.1, many of the datasets that exist at the time of writing suffer from a number of issues, including mislabelling of African American English and sampling issues (Wiegand et al., 2019). Similarly, many more recent datasets suffer from critical flaws, e.g. the dataset proposed by Salminen et al. (2018) which treats responses to physical abuse and the abuse of power as 'abusive'. The more recent datasets have not been subject to the same level of critical inquiry as the early datasets, in spite of many of similar issues being apparent upon inspection. As the limitations of early datasets are well known within the field, and their results can be interpreted within the context of their limitations, I find that these are most appropriate for use for the work in this thesis.

I also make use of two datasets that I have developed and published before starting my doctoral research, the *Hate Speech* and *Hate Expert* datasets. These were developed prior to undertaking my doctoral research and are historically significant as they a) help to map how my thinking has developed over time and b) were released as some of the first publicly available datasets for online abuse.

**Hate Speech** Published as the first publicly available dataset for hate speech and abusive language detection, Talat and Hovy (2016) developed a dataset for detecting abuse towards gendered and racialised minorities. In an interview in the Let's Chat Ethics Podcast, Zeerak Waseem shared that the initial motivation for developing the dataset was the somewhat naïve hope to address online harassment as experienced by women during #GamerGate, a harassment campaign against female game developers and games journalists (Massanari, 2015). This aim of developing tools that can protect marginalised people is apparent in the

[2] I only provide examples from datasets that are not annotated for abuse, as adding more than the minimum number of examples provides little additional value. Moreover, as all datasets are publicly available, examples are readily available for perusal for interested readers.

data sampling and the source of data. As a large amount of the GamerGate abuse occurred on Twitter, Talat and Hovy (2016) use Twitter as a source of their data, collecting $16,914$ tweets labelled as ‘sexist’, ‘racist’, and ‘neither’. In efforts to ensure that their collected and annotated sample contains gendered and racialised abuse, they bootstrap their corpus collection by first search for common slurs against women, ethnic minorities, religious minorities, and sexual minorities to identify the salient terms and users for scraping. To annotate this dataset, with the target group in mind, Talat and Hovy (2016) develop 11 questions to test whether a comment is hateful or not. This set of questions focuses on breadth in the types of hate expressed rather than depth in each type. This is apparent as the tests ranges from asking about explicit forms of hate, such as the use of slurs to implicit forms like questions around stereotyping and the use of straw man arguments in criticisms of minorities. Talat and Hovy (2016) annotate their dataset and have their annotations verified by an external annotator. Collectively, these decisions are made to ensure that there was a diversity in the forms of hate in addition to the sources. However, as they note, the racist abuse only comes from 9 different accounts. Moreover, as salient terms were sampled for annotation, some terms (i.e. the hashtag for the Australian TV show My Kitchen Rules) are over-represented in the data. In spite of these issues, the annotations in this dataset are *embodied* within the context of critical race and gender studies perspectives on abuse.

**Hate Expert** In an extension of the dataset proposed by Talat and Hovy (2016), Talat (2016) sample $6,909$ tweets from the original scrape and have it annotated by two groups, in efforts to understand the influence of annotator biases. The first group consisted of “feminist and anti-racism activists” (Talat, 2016) who annotate the sample of the dataset with one of four labels ‘racist’, ‘sexist’, ‘both’, and ‘neither’. The second group of annotators were recruited from CrowdFlower to re-annotate the sample annotated by the first group.[3] The label set was expanded by Talat (2016) to include the category ‘both’, in acknowledgement that marginalisation can be expressed across multiple dimensions, in an *Intersectional* manner. Comparing models optimised on each group of annotators, Talat (2016) find that models that use the annotations of the first group consistently outperform models optimised on the second. Talat (2016) argue that the reason for this difference is that the models that are optimised on the former group benefit from similarities in the understanding of hate speech. On the other hand, the only salient unifying characteristic for the latter is that they perform work on a micro-work platform, which produces internal inconsistencies in labelling that renders it harder for models to consistently identify patterns that they can optimise on.

[3] CrowdFlower has since been renamed Appen.

In presenting this dataset, Talat (2016) propose that ideologues can take similar positions on a topic, given their subjective positionalities. They argue that only through a principled understanding of hate speech is it possible to annotate reliably for hate speech and that crowd-sourced annotations for hate speech display inconsistencies that to some degree erases the meaning of the term. In using this dataset for this dissertation, I use the annotations provided by the feminist and anti-racist activists.

**Offence** Departing from the question of forms of hate speech and gender studies and critical race theory based annotation guidelines, Davidson et al. (2017) turn instead to ask where the distinction between simply 'offensive' speech and 'hate speech' lies. Using a list of terms from Hatebase to identify 33,458 users whose tweets they sample.[4] From these users, they randomly sample 25,000 tweets for annotation by CrowdFlower workers, resulting in 24,802 annotated tweets. The crowd-workers were given guidelines to aid them in distinguishing between 'offensive content', 'hateful content' or 'neither offensive or hateful', selecting only one for each tweet. Davidson et al. (2017) instruct their annotators that hate speech is speech that "is used in reference to certain groups that expresses hatred towards the group or is intended to be derogatory, to humiliate, or to insult the members of the group". Moreover, they provide examples of such content which includes the straightforward "Need to send these w******s back to their country"[5] and the more conflicted "I hate white trash". The conflict in the latter stems from it being unclear whether the emphasis is on class-based hate or if it is targeting white people. While the former is less contentious, the latter would imply that white people too are targets of marginalisation on the basis of their race. However, as numerous scholars have argued, whiteness is the hegemonic force that marginalises (Benjamin, 2019; Noble, 2018). 'Offensive' content is provided as an alternative, less serious form of potentially unwanted content. This group is defined in contrast to the hateful class "[o]ffensive content might use some of the same words we associate with hate speech but do not *necessarily* constitute hate speech because the words are not used in the same context as 'hate speech"'.[6] From this definition it's clear that in spite of the instruction to select only one category, Davidson et al. (2017) acknowledge that there is a potential overlap between offensive language and hate speech. Moreover, unlike the annotation guidelines proposed by Talat and Hovy (2016), the definition of offensive draws in the question of context. Illustrating this point, Davidson et al. (2017) provide the

[4] Hatebase.org is a website that crowd-sources slurs and insulting turns of phrase. Due to marginalised people being disproportionately targeted, there is a distributional skew towards terms that target marginalised people in the number of terms.

[5] Censoring of the slur is mine.

[6] Emphasis added.

following example "Oh shush you know I love you f****t". This use of context, provides space for inoffensive uses of slurs and insulting terms e.g. for reclaimed and in-group uses, a space that the annotation guidelines of Talat and Hovy (2016) does not afford. With the annotators being selected from CrowdFlower, the issues of multiple distinct ideologues in the annotator pool raised by Talat (2016) are likely also manifest in this dataset. However, as the dataset offers a space within which one can utter offensive but not hateful messages, it also offers the space to live, that is it offers spaces that dominant discourse on acceptability of language use would deem as unacceptable.[7] In consideration of the marginalisation of queer people and people of colour, this dataset thus offers space for their uninterrupted existence. However, as the dataset is labelled for a multi-class classification problem, where a single label is assigned to each document, the dataset does not afford space to be free from exposure to offensive language and hate speech, without also treating the two as equally sanctionable.

**Toxicity** Starting from a similar point as Davidson et al. (2017), Wulczyn et al. (2017) develop a dataset of $115,737$ comments to understand which types of conversations are likely to make users depart from the conversation. Departing from the early tradition of using Twitter as a source of data, Wulczyn et al. (2017) consider the Wikipedia editor discussion pages. Taking a narrow view of behaviours that inhibits participation in conversations, Wulczyn et al. (2017) focus on personal attacks and harassment, specifically asking their annotators whether which entity (the participant or a third party) is the subject of the attack. As a last positive category, they include whether it is "[a]nother kind of attack or harassment", thus relegating all forms of harassment that are not directed at specific individuals to a residual category. The dataset thus is comprised of 'personal attacks' and 'other forms of harassment'. As the study is specifically grounded in identifying personal attacks, this categorisation of various forms of personal attacks and a residual category as positive instances is in line with the aims of the data, if not the description. Using this definition, Wulczyn et al. (2017) select a random sample of $37,611$ comments that are labelled by 10 annotators for personal attacks. This resulted in only 0.9% of the labelled data in the positive class. To address this, Wulczyn et al. (2017) identify an additional $78,126$ comments that are sampled from users whose content had been moderated from the discussion pages. For each user, 5 comments made by the moderated user around the moderated comment were collected and subject to annotation, resulting in the positive class consisting of 16.9% of the total dataset. Similarly to Talat

[7] By dominant discourses on acceptability, I refer to what mainstream discourses deem as acceptable and unacceptable manners of speaking. However, such a discourses are internally inconsistent, as Dias Oliva et al. (2021) show, acceptable speech can come to include neo-nazi and white supremacist speech that threaten social cohesion while deeming speech by queer communities as toxic and inherently holding greater threat to the boundaries within which society should operate.

(2016) and Davidson et al. (2017), Wulczyn et al. (2017) use CrowdFlower to obtain their annotations and subsequently are prone to similar issues in their data. However, to curb such issues they obtain 10 annotations for each comment, allowing to compute a majority vote that takes a broader perspective on the comment into account. In spite of this approach, where those annotators are from and what their position on personal attacks are, and their ability to identify subtle attacks, still remain uncertain resulting in a dataset that may take a global position or a culturally grounded position on identifying personal attacks, e.g. if a large subset of annotators live in India, a subset of the data may very well reflect Indian perceptions of personal attacks. The resulting dataset has been constructed to understand which comments are likely to turn discussions "toxic" as a result of personal attacks. Through the use of 10 annotators for each comment, Wulczyn et al. (2017) aim for a global understanding of toxicity derived, in part, from personal attacks. Similarly to Davidson et al. (2017), there appear to be no consideration of the experiences of abuse against marginalised communities. Considering Wikipedia's well documented issues with being a hostile space to women (Torres, 2016) and the distribution of gender crowd-workers often veering towards a greater representation of men than women (Posch et al., 2018), the lack of such a consideration may further entrench subjective positions that are hostile towards women into the datasets and subsequently into the models.

**StormFront** Focusing on the white supremacist web-forum StormFront, Garcia et al. (2019) collect a dataset of $10,568$ sentences annotated by three of the authors for containing hateful utterances. Similarly to Davidson et al. (2017) and Wulczyn et al. (2016), Garcia et al. (2019) employ a marginalisation-blind definition and understanding of hate speech. In the case of a white supremacist web-forum, employing a marginalisation-blind definition is unlikely to be challenged as the participants are unlikely to engage in derogation against white, straight, cisgender men. The decision for using StormFront as a source of data was motivated by the prevalence of "pseudo-rational discussions of race". Moreover, this dataset further distinguishes itself from prior datasets by annotating on a sentence level. The authors argue that annotating on a sentence level can reduce the confounding factors by only addressing content which is explicitly hateful. While this may, in some instances have little effect as the surrounding sentences bear no impact on whether a sentence is hateful. This particularly holds for explicit hate speech. However, for subtle forms of hate speech, conducting sentence level annotation may obscure hate that is only apparent when considering a post in its entirety rather than its sentence level components. In order to address this issue, Garcia et al. (2019) introduce a 'related' tag which is to be used when individual sentences do not convey hate but the combination of several sentences in sequence do convey hate. This method for mitigation does not account for longer sequences of sentences that convey hate, as is often the case for subtle forms of hate speech and dog whistles. Moreover, as Garcia et al. (2019) take a very conservative position on what constitutes hate, for instance, the use of a derogatory term, on a white supremacist web-forum, "cannot be said to be a deliberate attack, taken without any more context, despite it likely being offensive." For this reason, Garcia et al. (2019) argue that simply the occurrence of slurs weaponised against marginalised communities cannot be said to be hateful. Thus, while initially side-stepping the issue of marginalisation-blind definitions by sourcing data from a white supremacist web-forum, it is softly reintroduced by taking a conservative stance on what constitutes hate.

#### 3.1.2.2 Non-abuse Datasets

**Sarcasm** Oraby et al. (2016) develop a dataset for sarcasm detection in dialogues. The dataset was developed in order to address the lack annotation for subtypes of sarcasm, i.e. rhetorical questions and hyperbole, at scale in previous datasets. Sourcing their data from the Internet Argument Corpus (IAC) (Abbott et al., 2016), Oraby et al. (2016) annotate their data for "generic sarcasm, rhetorical questions, and hyperbole". In order to generate a dataset from the IAC, Oraby et al. (2016) optimise a 'weakly-supervised pattern learner' (Oraby et al., 2016) to identify a set of $30,000$ posts, filtering two thirds of the posts that don't contain any 'not-sarcastic' cues and annotate the remaining $11,040$ posts in quote-response pairs for annotation on Amazon Mechanical Turk. Similarly to the abusive language datasets annotated on CrowdFlower, this choice of annotators can introduce biases stemming from the subjective embodiments of the human annotators and the geo-cultural contexts in which they exist. Following the annotation process a dataset of $6,520$ posts (with a 50% split of sarcastic and not-sarcastic posts) is obtaned and released. Examining the dataset for suitability for machine learning experiments, Oraby et al. (2016) optimise a linear SVM with Stochastic Gradient Descent (SGD) optimisation and L2 regularisation obtaining *F1-score* of 0.74 using features derived from Word2Vec (Mikolov et al., 2010).

**Argument Basis** Investigating the characteristics of factual and emotional argumentation styles, Oraby et al. (2015) also draw on the IAC as the source of data. Considering quote-response pairs, each response is annotated for whether the argument presented in the response based primarily in fact or feeling. Oraby et al. (2015) present $10,003$ from the IAC for annotation by $5-7$ crowd-workers on Amazon Mechanical Turk for annotation selecting a value ranging from $-5$ to $5$ to indicate whether the response is a feeling or fact-based argument, where negative values indicate that the argument basis is dominated by an emotional argumentation style and positive values indicate a fact-based argument. Each

| Document | Label |
|---|---|
| “But what about the married couple who know right up front they do not intend to have children, don’t even intend to try? Do we ban them from marriage to?” | Not Sarcasm |
| “wow! i’ve never thought about it like that. you’ve given that a lot of thought ;)” | Sarcasm |
| “She had plenty of choice in the matter. She should have chosen to keep the kid. Even if she lost all her money because of this it would have been the right thing to do” | Feeling-based Argument |
| “Mutation is not always via radiation.. there also can be inaccurate duplication of dna. As for new species happening, it has been observed in nature.” | Fact-based Argument |
| “Holy muther Sandy is no joke Sending the love and good vibes” | Care |
| “Your desire to make sacrifices for humanity are in vain if you’re going to kill human beings in the proces” | Harm |
| “Methinks Obama hearts Sandy Gives him plausible excuse to avoid all questions related to Benghazi” | Subversion |
| “#BlackLivesMatter of course they do but so do #AllLivesMatter. Protests are silly and simply bully tactics. Respect police and get respect” | Authority |
| “I’ll follow any fellow servicemen and loyal AT_USER supporters” | Loyalty |
| “#AllLivesDidntMatter when ppl failed to show support & empathy by denigrating #BlackLivesMatter to #AllLivesMatter” | Betrayal |
| “#PrayersForBaltimore #AllLivesMatter #StopTheViolence #FreddieGray and all the other families deserve justice #NoJusticeNoPeace” | Fairness |
| “People who take advantage of crises like Sandy to harm deceive or otherwise be a dick to those in need you re going to THE SPECIAL HELL” | Cheating |
| “Prayer is the only means of bringing about orderliness and peace and repose in our daily lives” | Purity |
| “Everyone should unfollow AT_USER immediately Making hurricane jokes is pathetic and insensitive and disgusting sandy” | Degradation |
| “5 Leadership Priorities During Times of Crises #BaltimoreRiots #leadership” | Non-Moral |

**Table 3.2** Examples of documents in each class of the *Sarcasm, Argument Basis,* and *Moral Sentiments* datasets.

document is then given a binary label indicating its argument basis, where all texts with a score greater than 1 are assigned as fact-based, all texts with a score lower than $-1$ are assigned to the feeling-based class, and all scores $[-1,1]$ are discarded. This annotation process results in $3,466$ fact-based and $2,382$ feeling-based documents. Similarly to the previously examined datasets that utilise crowd-workers, this dataset is also subject to the contexts which the individual annotators exist within. For instance if an annotator is from a culture where feeling-based argumentation is not experienced as impassioned but instead supportive of facts, they may be likely to rate some documents as more fact-based than annotators who hail from cultures that emphasise fact-based argumentation would deem as relying on an emotional argumentation style. The subjectivity of the annotation task may provide an explanation for why $4,155$ or more than $41\%$ of the documents are discarded due to being rated, in aggregate, as dominated by neither fact or emotion.

**Moral Sentiment** The final dataset used in this dissertation is the Moral Foundations Twitter Corpus (Hoover et al., 2019). This dataset provides $35,108$ tweets annotated for 10 different categories of moral sentiment, introducing the task of moral sentiment prediction. A task, and dataset designed to allow psychology researchers to investigate the relationship between comments made around events and the moral foundations found in such comments made on Twitter. Hoover et al. (2019) draw from research in psychology around human morality using a five-factor taxonomy that reveals insights into the moral foundations that underlie comments about and attitudes towards topics. Each of the five factors are represented through a binary, where one end of the binary represents a virtue and is contrarian to the other, representing a vice. Hoover et al. (2019) argue that the human expression of vice and virtue are distinguishable from one another through distinct language use for each. The five factors introduced are `care`, "concerns related to caring for others" and `harm`, "concerns related to not harming others"; `fairness`, "concerns related to fairness and equality" and `cheating`, "concerns related not not cheating or exploiting others"; `loyalty`, "concerns related to prioritising one's ingroup" and `betrayal`, "concerns related to not betraying or abandoning one's ingroup"; `authority`, "concerns related to submitting to authority and tradition" and `subversion`, "concerns related to not subverting authority or tradition"; `purity`, "concerns related to maintaining the purity of sacred entities, such as the body or a relic" and `degradation`, "concerns focused on the contamination of such [sacred] entities." Noting that there is low occurrence of moral sentiments expressed in a random sample of tweets, Hoover et al. (2019) collect tweets related seven different discourse domains where the occurrence of moral sentiment is likely to at a high rate: Black Lives Matter, All Lives Matter, Baltimore protests following the death of Freddie Gray, the 2016 presidential elections in the United States of America, hurricane Sandy, the #MeToo movement, and offensive language, re-annotating a sample of Davidson et al. (2017) for the moral sentiments. For annotation, Hoover et al. (2019) train 8 undergraduate research assistants to an expert-level familiarity with the moral foundations taxonomy, annotating $4,000-6,000$ for each discourse domain. The annotators are trained through training sessions and, in early stages, discussion surrounding annotator disagreement. The annotator selection procedure here thus develops on the suggestion of Talat (2016) to use expert annotators to describing a means of training expert annotators for a highly subjective task. Interestingly, as the annotation process continues past early stages, annotator disagreements are not resolved, instead the authors opt for expressing the inherent subjectivity of the human annotation task.

#### 3.1.2.3 Non-English Datasets for Abuse

In this dissertation, I focus my attention to detecting abuse in English language datasets as my methods do not map to other languages. However, an important growing body of research and resources are being developed for other languages such as Arabic (Albadi et al., 2018; Mulki et al., 2019; Ousidhoum et al., 2019), Croatian (Ljubešić et al., 2018), Danish (Sigurbergsson and Derczynski, 2019), French (Chiril et al., 2019a; Chung et al., 2019), and Urdu (Rizwan et al., 2020) amongst many more.

Developing models for each individual language, and in particular resources that address abuse that code-switches, require an attention to the particularities of the different languages and cultures, just as model development for English requires researchers to be attuned to the particularities and cultures represented in English language use.

### 3.1.3 Generalisable Machine Learning Models for Abusive Language Detection

A common criticism of many current computational methods for abuse detection is that they have poor generalisability onto other datasets. Although this issue of non-generalisability poses a serious issue for the abuse community, it has received relatively little attention (Fortuna et al., 2021; Karan and Šnajder, 2018; Swamy et al., 2019; Talat, 2016; Talat et al., 2018; Wiegand et al., 2019) in comparison to single-dataset classifier performance. In each of the computational chapters (see chapters 5 and 6), I also provide consideration of how well the optimised models perform on out-of-domain datasets. In the pursuit of models that generalise well onto other datasets, researchers have proposed a variety of architectures. As an initial investigation into the question of generalisability, Talat (2016) note that the best performing classifier on the dataset they propose does not generalise well onto the *Hate*

*Speech* dataset, noting that the performance of their classifier drops by more than a 25%. Using MTL,[8] Talat et al. (2018) address the issue of poor generalisability between the *Hate Speech* and *Hate Expert* (combined into a single dataset) and the *Offence* dataset, showing that a MTL framework can be used for optimising models that can generalise onto from one cultural context onto another. Moreover, considering the results posted by Talat et al. (2018), it appears that there is a trade-off between well-performing in-domain models and well-performing cross-domain models, where cross-domain improvements appear to come at the cost of in-domain performance, where out-of-domain performance is computed by mapping the classes in the in-domain datasets to the out-of-domain dataset.

Karan and Šnajder (2018) further explore the question of cross-dataset generalisability using a linear SVM model. Karan and Šnajder (2018) approach the task of out-of-dataset performance as a classical domain adaptation task, finding that without significant procedures for domain adaptation, there is poor generalisability. Similarly to Talat et al. (2018), Karan and Šnajder (2018) find that cross-domain performance comes at the cost of in-domain performances but with large out-of-domain improvements. One difference between Karan and Šnajder (2018) and the previously described studies is that Karan and Šnajder (2018) reduce the learning task to a binary classification task of 'abusive' and 'non-abusive' documents.

The last approach to generalisation I consider is the work by Fortuna et al. (2021). In this paper, the authors compare four different models for out-of-domain classification: a Bag-of-Words SVM model, a Continuous Bag-of-Words FastText model, a BERT model (Devlin et al., 2019), and an ALBERT model Lan et al. (2020). The latter two being transformer-based language-models that are fined-tuned to the task of predicting abuse. Fortuna et al. (2021) propose a different class organisation to past studies, first they propose as generalised class organisation that collapse classes across datasets into a smaller, generalised subset that maps across datasets. For instance, the 'sexist' class provided by Talat and Hovy (2016) and the 'misogyny' class provided by Fersini et al. (2018) into a 'misogyny-sexism' class. Each of the generalised classes are binarised to allow models optimised with other standardised labels to predict on them. Using these generalised classes, Fortuna et al. (2021) show that by using methods that capture more complex word-interactions, out-of-domain performance generally improves within and out of domain, subject to the classification task. Specifically, they find that when classes have significant overlaps across datasets in their rationalisation of what the are to represent then models optimised on those classes will map well onto the rest. Conversely, when the classes have a little overlap, the models will generalise poorly onto the new dataset. Moreover, Fortuna et al. (2021) identify that some dataset combinations produce

[8] Multi-Task Learning allows for optimising models using multiple different datasets, for distinct machine learning tasks, where one (main) task is given higher priority and all other tasks are treated as auxiliary tasks.

poor generalisation between each other regardless of the models used. This, in concert with their conclusion that dataset overlap and out-of-domain similarities are drivers of model generalisation has two implications. First, current computational models can, to some degree, adapt onto new distributions and samples but models using words as input are poorly suited for learning general trends of a wide variety of abuse, including closely related concepts such as 'toxicity' and 'severe toxicity' (Fortuna et al., 2021). Second, as models do not generalise onto other concepts, even if closely related, research in the detection of online abuse must either develop methods that can generalise onto studying different objects and perspectives of online abuse, or datasets must be annotated following highly similar annotation guidelines at the cost of the depth and breadth of concepts that can be explored.

## 3.2 Modelling Techniques

In this section I provide an introduction to the different modelling techniques that I use throughout the dissertation.

### 3.2.1 Data Encoding

In order for models to read the data, it is necessary to provide the models with machine readable representations of the data. The first step to creating such machine readable representations is to provide each unique token with a numerical index. The numerical index, and what it represents is a matter of how the data is pre-processed. For instance, in chapter 5, I represent tokens in three different ways. First, I represent tokens using their surface form, that is each word is represented in its entirety following a tokenisation process where all words are lower-cased and punctuation markers are split from the word (see chapter 5 for further pre-processing steps). Second, I take the surface forms of tokens computed and represent them as the categories of the Linguistic Inquiry and Word Count (LIWC) categories each token induces (see chapter 5 for further detail). Finally, in chapters 5 and 6 I represent tokens as the subwords that they consist of. In this section, the sub-word forms while omitting the surface-token and LIWC-token forms as these rely on simple pre-processing and mapping steps that are described in more detail in chapter 5.

#### 3.2.1.1 Byte-Pair Embeddings

Byte-Pair Encodings were introduced to the NLP comunity by Sennrich et al. (2016) for the task of Neural Machine Translation to address the issue of out-of-vocabulary tokens. In this paper, the authors argue that for word-level machine translation there is not always a one-to-one relationship between a word in the source language and its translation into a target language. Sennrich et al. (2016) illustrate this point through compound words, where a compound word represents a specific entity that is represented through multiple words in the source language, e.g. the German *Abwasser|behandlungs|anlange* and its English translation *sewage water treatment plant* (Sennrich et al., 2016). Sennrich et al. (2016) propose to compute sub-words using the byte-pair encodings algorithm proposed by Gage (1994). While the algorithm proposed by Gage (1994) operates on bytes and seeks to develop a new representation of bytes that can compress their representation, Sennrich et al. (2016) seek to operate on the sub-units of words, that is a sequence of characters. In both cases, the algorithm operates by considering the input and identifying frequently occurring patterns that can be represented in terms of a single unit. In efforts to obtain a sub-word representation, Sennrich et al. (2016) initialise their algorithm initially with a vocabulary consisting of each unique character token in the dataset and then count all symbol pairs (e.g. character co-occurrences) and merge the most frequently occurring pairs and adding it to the vocabulary. This merging process is repeated a number of times, where the total number of merge operations is a hyper-parameter set by the designer of the sub-word representation. The size of the vocabulary following this process will be the size of the original vocabulary plus the number of merge operations that are set by the designers (Sennrich et al., 2016) In terms of language, using sub-words to represent documents can minimise the number of out-of-vocabulary tokens in the validation and evaluation sets of a dataset, as the likelihood of a word not being represented decreases as it is broken down into its subwords.

In this dissertation, I use the pre-optimised Byte-Pair Embeddings (BPE) developed by Heinzerling and Strube (2018). These embeddings were optimised for 275 languages using the Wikipedia pages in each language as the source of data. Heinzerling and Strube (2018) provide embeddings for $1,000$, $3,000$, $5,000$, $10,000$, $25,000$, $50,000$, $100,000$, and $200,000$ merge operations with dimensions 25, 50, 100, 200, and 300. For all chapters in this dissertation, I choose the 300 dimensional embeddings that have been subject to $200,000$ merge operations as these embeddings are likely to offer good representations of the data that used in this dissertation.

### 3.2.2 Strategies Against Over-fitting

Machine learning models are prone to identify salient patterns in the optimisation data with the result that they perform poorly on evaluation sets and out-of-domain data. In order to address this issue, I use a number of different techniques depending on the type of model used. For linear models, I experiment with three different regularisers: L1 regularisation, L2 regularisation, and Elasticnet. For neural networks, that by virtue of their ability to identify and represent complex interaction patterns are prone to overfit, I use two different techniques, namely early stopping and dropout.

**L1 Regularisation** L1 regularisation operates by iteratively zeroing out uninformative features in order to produce a more sparse representation of the data while minimising loss of performance of a given model (Goldberg, 2017). For instance, if there are two features $x_1$ and $x_2$ that both carry an equal weight towards the same class, one of the features will be zeroed out while the other will retain its weight. While this can be helpful in an in-domain setting, it may not be quite as useful when the model is used on new data, in cases where $x_1$ exists in the document to be classified but is zeroed out $x_2$ does not occur in the document.

**L2 Regularisation** To address this short-coming, $L2$ regularisation is proposed (Goldberg, 2017). $L2$ regularisation seeks to penalise weights of features, making the weights smaller, rather than altogether zeroing out any weights. This penalisation and reduction of weights by $L2$ regularisation seeks to minimise across all features. Thus $L2$ regularisation does not necessarily zero out any individual feature but instead reduce the weight of all features to prevent over-fitting to any particular set of features.

**Elastic Net** Elastic Net seeks to combine $L1$ and $L2$ regularisation into a single regularisation function (Goldberg, 2017). Thus, elastic net seeks to both zero out uninformative features and minimise the weight of all features to reduce variance between them. Elastic Net is particularly fitting in modelling contexts where there is a high dimensionality in the data, for which reason it is desirable to reduce the size of the feature space while retaining a maximum number of features that are informative towards the prediction task.

**Dropout** In order to prevent neural networks from over-fitting on optimisation data, Goldberg (2017) introduce the notion of dropout. Dropout refers to randomly zeroing out some values of a model's internal representation between different layers. The idea behind dropout is that models may over-fit to individual tokens or interaction patterns between tokens, thus to prevent the model from learning such patterns, one can randomly zero out values in the internal representations of a document as it is passed through the layers of the model. Such zeroing out forces the model to adapt to different representations for a given document each time it is passed through the model and, hopefully, identify general patterns rather than ones that occur from spurious correlations in the data.

**Early Stopping** A second method for addressing over-fitting in neural network models is the idea of early stopping (Prechelt, 1998). Early stopping, in terms of neural networks, means to end a optimisation cycle before it passed over the data for the number of epochs specified by the researcher. The idea behind early stopping is that a model may identify an optimal representation before the maximum number of epochs has been reached. Any further optimisation processes on the model representation are thus likely to have a detrimental effect to a model's performance on the evaluation set. In this dissertation I trigger early stopping by considering the development of model loss. Specifically, if the model loss monotonically increases for a set number of epochs, I trigger early stopping as this indicates that the model has already identified a representation that minimises the loss.

### 3.2.3 Optimisation Techniques

Optimising a neural network requires a host of different techniques for optimisation, such that the model can identify optimal minima. Among these are the loss function, the activation function and pooling functions for CNNs. Further, in order to identify optimal minima, it may be necessary to optimise the model with a number of different values for the hyper-parameters (i.e. model parameters such as embedding sizes and parameters for the optimisation functions such as the learning rate) which is also a process that can be subject to optimisation itself. Here, I describe the different optimisation techniques that I use in this dissertation.

Across all neural network models optimised for the experiments in this dissertation, I use a `softmax` function to produce output values representing the likelihood for each class based on the model representations. The `softmax` function operates by taking a vector and producing a value of $[0,1]$ of the vector by computing the normalised exponential function of all the units in the layers (see eq. (3.1) for a mathematical definition for `softmax`).

$$S(\mathbf{x})_i = \frac{e^{\mathbf{x}_j}}{\Sigma_j e^{\mathbf{x}_j}} \tag{3.1}$$

**Fig. 3.1** Equation for the `softmax` function (Goldberg, 2017).

Moreover, as the resulting vector sums to 1 we can treat the values in the vector as a probability distribution where the largest value represents the most likely class.

#### 3.2.3.1 Loss

Broadly, two different types of neural networks exist: feed-forward networks and networks that use back-propagation. Feed-forward networks chronologically update the model on the basis of the data it is provided without concern for how each update to the model's parameters impact the model's ability to perform the classification task. Back-propagation was introduced as a method with which model parameters could be updated after the forward step of the model had completed and an evaluation of the model's performance with its most recent parameter weight s(Goldberg, 2017) . By obtaining the model's loss, or model error given by a loss function, one can back-propagate the loss through the model to perform an update to the model's parameters after the completion of the forward step. In this dissertation, I only make use of back-propagated models with `Negative Log Likelihood` loss.

**Negative Log Likelihood** I choose `Negative Log Likelihood (NLL)` as a loss function as it is particularly well-suited for use with the `softmax` function. I provide the definition of `NLL` as provided by the PyTorch library (Paszke et al., 2019) (see eq. (3.2)), where $\hat{\mathbf{y}}$ is a vector of the predicted labels, $\mathbf{y}$ is the vector of given labels (Goldberg, 2017).[9] Negative log-likelihood operates by by assigning a higher loss to for the correct class for each document on the basis of the probability estimates (obtained through the `softmax` function) for the class. The higher the probability estimation for the correct class is, the lower the loss is and on the other hand the smaller the probability estimate for the correct class is, the higher the loss is. By focusing on the probability estimate for the correct, in terms of ground truth, label, `NLL` avoids the potential issue of assigning all predictions with a correct or incorrect label a specific value. Thus, `NLL` addresses the model's certainty rather than the prediction itself.

$$L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{i} \mathbf{y}_i \log(\hat{\mathbf{y}}_i) \tag{3.2}$$

**Fig. 3.2** Equation for `Negative Log Likelihood` loss.

#### 3.2.3.2 Non-linearities

In the experiments conducted in this dissertation I use two different non-linear functions that I subject various layers in the neural network models to. The role of non-linearities in neural networks is to allow for models to optimise non-linear functions rather than linear ones. The two non-linearities that I experiment with are `Tanh` and `ReLU`.

The `hyperbolic tangent`, or `Tanh`, function (see eq. (3.3) for its mathematical definition) is a non-linear activation function that element wise transforms the values of the tensor representation of the model into a real valued space between $[-1,1]$. `Tanh` is a monotonically

[9] See `https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html` for the implementation details for `Negative Log Likelihood` used in PyTorch.

increasing function that is symmetrical around 0 due to which there is a risk of the issue of vanishing gradients for the model (Teuwen and Moriakov, 2020). Vanishing gradients refers to the issue where the gradients of the models become increasingly small, to the point of no longer having an effect on the parameter updates, due to being centred around 0.

$$\tanh x = \frac{e^{2x} - -1}{e^{2x} + 1} \tag{3.3}$$

**Fig. 3.3** Equation for `Tanh` (Goldberg, 2017).

One way to address the potential issue of vanishing gradients is to use a `Rectified Linear Unit` (`ReLU`) as the activation function (see eq. (3.4) for the mathematical definition of `ReLU`). Unlike the `Tanh` function, `ReLU` is not a symmetrical function, but instead relies on a binary evaluation of each element in a vector. If the weight of the element under consideration $w_x < 0$, then the value computed is $ReLU(x) = 0$. On the other hand, when the value $w_x > 0$, then the value computed is $ReLU(x) = 1 \cdot w_x$ (Teuwen and Moriakov, 2020).

$$ReLU(x) = \max(0, x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases} \tag{3.4}$$

**Fig. 3.4** Equation for `ReLU`.

#### 3.2.3.3 Pooling Layers

For CNN models it is necessary to use either average pooling or maximum pooling to summarise the features under a filter. As a summary, pooling operations act as a method for downsampling the feature representation obtained after convolutional layers. Two common kinds of pooling operations are average pooling and maximum pooling. Average pooling computes the mean value of the pooling features under the filter while maximum pooling extracts the largest value. In my experiments with CNN models, I use maximum pooling exclusively due to its dominance in the CNN models developed for abuse detection (Kolhatkar et al., 2020).

#### 3.2.3.4 Optimisation Algorithms

At the heart of neural networks lie the optimisation algorithms that control the rate and manner in which model weights are updated. A number of different optimisation algorithms have been proposed for neural networks, but in my experiments I focus on two algorithms, Stochastic Gradient Descent (SGD) and Adam. For each of these algorithms, I use the originally proposed algorithms, SGD and Adam, and a variant that address specific short-comings of each algorithms, Averaged Stochastic Gradient Descent (ASGD) and Adam with decoupled weight decay (AdamW). For all algorithms, we use the implementations used in Paszke et al. (2019) and refer readers to the PyTorch documentation for further details.[10]

**Stochastic Gradient Descent** Stochastic Gradient Descent (Sutskever et al., 2013) is a popular optimisation algorithm used for neural networks and has been used by a large number of researchers for a diverse set of tasks across various machine learning research areas and tasks, including online abuse detection (Bodapati et al., 2019; Singh et al., 2018). SGD relies on gradient descent, which is an algorithm that computes the gradients of points on a function until it reaches a minima. This process can be computationally expensive as gradient descent requires a computation on the entire dataset. This approach has two issues: First it is computationally expensive as the computation is performed on the entire dataset; second, gradient descent requires a learning rate being given, which determines the position for the next point at which to compute the gradient. If the learning rate is sufficiently small, and the function under optimisation is not a convex function, gradient descent may identify and settle at a local minima rather than the desired global minima.

$$v_{t+1} = \mu v_t - \varepsilon \nabla f(\theta_t + \mu v_t) \tag{3.5}$$

$$\theta_{t+1} = \theta_t + v_{t+1} \tag{3.6}$$

**Fig. 3.5** Equation for Stochastic Gradient Descent (Sutskever et al., 2014), where $\varepsilon$ is the learning rate, $\nabla f(\theta_t + \mu v_t)$ is the gradient, and $\mu$ is the momentum.

SGD similarly computes the gradients of points on a function, but rather than computing the gradient descent on the entire dataset, a single example is selected and the gradient descent is computed for that data point, thus minimising the computation time, even when more iterations are necessary to identify the minima. The second issue of local minima is in part addressed by the random nature of selecting a data point to compute the gradient from. This randomness results in greater fluctuations in the development of the gradient, however, this exact fluctuation and variance may allow the algorithm to identify a better minima.

[10] The API reference can be found at https://pytorch.org/docs/stable/optim.html.

**Averaged Stochastic Gradient Descent** Another approach to addressing the issue of identifying optimal minima is `Averaged Stochastic Gradient Descent` Polyak and Juditsky (1992). `ASGD` operates similarly to `SGD` but considers averaged trajectories in order to accelerate the identification of the optimal minima. The acceleration is obtained through a reduction of noise from the stochastic nature of the selection of data point for consideration. `ASGD` takes the standard `SGD` algorithm and recursively computes the average $\bar{w_t} = \frac{1}{t}\Sigma_{i=1}^{t} w_t$ (Bottou, 2010).

**Adam** The `Adam` algorithm (Kingma and Ba, 2015) is also frequently used in abuse detection classification research (Kolhatkar et al., 2020; Meyer and Gambäck, 2019; Zimmerman et al., 2018). The algorithm seeks to further push the goal of faster convergence onto optimal minima, Kingma and Ba (2015) propose the `Adam` algorithm. The `Adam` algorithm is also a stochastic optimisation algorithm, however it only requires computing the first-order gradients. The algorithm seeks to compute the value of parameters $\theta$ at time-step $t$ that achieves convergence. However, rather than updating all parameters with with the same learning rate, `Adam` maintains a learning rate for each parameter which is adapted as the optimisation of the network proceeds.

`Adam` achieves this by first computing the gradients with regard to the stochastic objective at time-step $t$, then updating the biased mean estimate and the biased uncentred variance estimate. This is followed by computing the bias-corrected mean and uncentred variance estimates, respectively which are computed by factoring in exponential decay rates for the moment (mean and uncentred variances) estimates. Finally, the value of $\theta_t$ is updated and the process is repeated if $\theta_t$ has not converged.

**Adam with Decoupled Weight Decay** `Adam with decoupled weight decay (AdamW)` was proposed by Loshchilov and Hutter (2019) as a result of examining the implementation of Adam in many libraries for neural network optimisation and finding that many had incorrectly implemented `Adam` using $L2$ regularisation rather than weight decay.

Since this correction, papers on abuse detection have started to use this algorithm (Röttger et al., 2020; Vidgen et al., 2020a) over the initial `Adam` implementation that used $L2$ regularisation rather than weight decay.

### 3.2.4 Bayesian Hyper Parameter Tuning

The performance of neural network architectures rely on a range of hyper-parameters that control their behaviour from a number of different positions in the model. For instance, the size of the layers in the network can be treated as a hyper-parameter, the learning rate for the optimisation algorithms, and the rate with which to apply dropout in the model. As a result of the many different potential hyper-parameters that can be tuned, the complete search space for all hyper-parameters grows exponentially for each new hyper-parameter under consideration. While the same holds true for linear models, the number of parameters to be explored often figure in much smaller ranges. For instance, the searches for parameters and hyper-parameters, for the linear models used in this dissertation are concluded in only minutes due to a smaller search space. For neural network models, a full search of the hyper-parameter search space however quickly becomes infeasible as the number of parameters to be explored scale. This introduces the question of how a hyper-parameter search space can be adequately explored without the need for a complete search through every possible combination.

One way to perform a hyper-parameter space search, without searching the complete space of every combination, is through Bayesian Optimisation for hyper-parameter identification (Snoek et al., 2012). Through the use of Gaussian Processes (GP), the selection of hyper-parameters for trial can be cast as an optimisation problem, where the hyper-parameters serve as a feature space and the performance obtained with each parameter serves as the label. The aim of the GP model is to estimate how each hyper-parameter contributes to the final classification performance of the model and provide suggestions for the next set of hyper-parameters to trial. I use Biewald (2020), which implements Snoek et al. (2012), for all hyper-parameter searches for neural network-based experiments.

### 3.2.5 Metrics

Model performances can be evaluated in a number of different ways, from qualitative analyses of the model outputs to quantitative analyses. Within the bracket of quantitative analysis, further subdivisions exist including the one I will use in this dissertation, namely the use metrics computed using model predictions and the ground truth. For my evaluation, I use the `F1-score`, `precision`, `recall`, and `accuracy`. As many of the datasets that are used for optimisation and evaluation have heavily imbalanced class distributions, each of these scores provide different aspects into model performances and different levels of insight into the models. The metrics all require insights into the agreements between the ground truth and a model's predictions. These agreements can be categorised into four different groups: `True Positives (TP)`, where the model's prediction and the ground truth label agree and the label belongs to the positive class; `True Negatives (TN)`, similarly where model prediction and ground truth agree and the label belongs to the negative class; `False`

`Positive (FP)`, where the model predicts the label for the positive class but the ground truth is in the negative class; and `False Negative (FN)`, which is the inverse of `True Positive`, i.e. the model predicts the negative class but the ground truth is in the positive class. [11]

**Accuracy** `Accuracy` is the simplest metrics among those I use, and it's subsequently also highly volatile to class imbalances. The score (see eq. (3.7)) computes the number of correct predictions out of all predictions made. For balanced datasets, this metric provides a good insight into a model's overall performance, however for imbalanced data, it is susceptible to providing a distorted view of a model's performance. For instance, if a dataset has a heavy class imbalance, a model that only predicts the majority class will have a deceivingly high `accuracy` score.

$$accuracy(Y,\hat{Y}) = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.7}$$

**Fig. 3.6** Equation for the *accuracy score*.

**Precision** `Precision` provides an estimate of how well a model predicts into the positive class. Specifically *precision* asks to which degree classifications into positive class are correct classifications into the class (see eq. (3.8)). Thus, one can ascertain to which degree a model can be trusted when it predicts a positive label.

$$precision(Y,\hat{Y}) = \frac{TP}{TP+FP} \tag{3.8}$$

**Fig. 3.7** Equation for the *precision score*.

**Recall** `Recall` on the other hand, provides insight into the ability of a model to retrieve correct instances of the positive class. By computing the fraction of data predicted correctly into the positive class and the union of data correctly predicted into the positive class or incorrectly predicted into the negative class, `recall` can allow for an intuition into how trust-worthy a model is when it predicts that data is not in the positive class.

[11] I do not use the `Area Under the receiver operating characteristic Curve` as a metric as this metric does assumes a similarity between all underlying samples (Stevenson, 2021). A guarantee that cannot be made when datasets with distinct annotation strategies and sampling are involved .

$$recall(Y,\hat{Y}) = \frac{TP}{TP+FN} \tag{3.9}$$

**Fig. 3.8** Equation for the *recall score*.

**F1-score** In practice it is often desirable to balance `precision` and `recall` as as they allow for intuitions into two crucial aspects of model performance, it's ability to correctly retrieve data into and exclude data from the positive class. The `F1-score` provides exactly such a balancing by computing the harmonic mean of the `precision` and `recall` scores (see eq. (3.10)).

$$F1\text{-}score(Y,\hat{Y}) = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.10}$$

**Fig. 3.9** Equation for the *F1-score*.

For abuse detection, the macro average of the *F1-score* is often used. The macro averaged *F1-score*, or macro *F1-score*, sum the *F1-score* for each class and computes their mean, thus providing insight into the performance of models across the different classes. In this dissertation, all *F1-scores* reported are macro *F1-scores* as this has been widely used in the abusive language detection field (Fortuna et al., 2021).

### 3.2.6 Machine Learning Models

As I explore different experimental research questions, I optimise different machine learning algorithm for detecting abusive language. Each of the model types that I use rely on different methods of operationalising data to obtain internal representations of the different classes. Here $X$ is to mean the processed input to the model, $Y$ is to denote the corresponding ground truth labels, and $\hat{Y}$ denote the set of model predictions. For all models, the aim is to optimise a function $f(X|Y)$ that can delineate between each class $y_i \in Y$.

**Logistic Regression** The first linear model that I use in this dissertation is Logistic Regression (LR), which has previously been used widely in NLP tasks. Logistic Regression is a model that carries certain assumptions about the data that is represented, in particular, I call to attention its assumption of feature independence. The assumption of feature independence presumes that each individual feature, or word token in the case of language, contributes to the classes in isolation of all other features. Trivially, this assumption does not hold for language.

In terms of mathematical modelling, LR relies on the `Sigmoid` function (see Equation 3.11) which calculates the probability of a data point, or document, belonging to a class. In Equation 3.11, $w_0, w_1, \ldots, w_n$ denote model coefficients that are obtained through maximum likelihood estimation and $x_1, \ldots, x_n$ represent the features that are treated independently.

$$F(\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}}} \tag{3.11}$$

**Fig. 3.10** The `sigmoid` function.

**Support Vector Machines** The Support Vector Machine (SVM) algorithm seeks to identify a hyper-plane where the data can be mapped to and classes $y_i \in Y$ are linearly separable. Such mappings can be computed using different kernels. Beyond identifying a hyper-plane where the classes are linearly separable, SVMs also have the additional aim of identifying a hyper-plane that maximises the margins, that is the distance between the linear separation and the closest data points for each class. Specifically, SVMs seeks to maximise the prediction given by $sign(w^T\phi(x)+b)$ where $\phi$ is the identity function and $b$ is an independent value. Although many different kernels exist for SVM classifiers, I use a linear kernel (see eq. (3.12) for the mathematical formula used by Pedregosa et al. (2011)) as this provides weights for each feature that can be analysed to understand which patterns the model is learning.

$$\min_{w,b} \frac{1}{2} w^T w + C \cdot \Sigma_{i=1} \max(0, y_i(w^T\phi(x_i)+b)) \tag{3.12}$$

**Fig. 3.11** Formulation of the Linear Support Vector Machine provided by Pedregosa et al. (2011), where $\phi$ is the identity function and $C$ is the regularisation strength.

**Multi-Layered Perceptron** The Multi-Layered Perceptron (MLP) is perhaps the simplest form of neural network. This neural network is an extension of the Perceptron algorithm by chaining several Perceptron units into a single layer, A Perceptron is updated given the update rule in Equation 3.13. Moreover, rather than consisting of a single Perceptron that learns weights of the optimisation data, MLPs are formed of multiple layers of chained Perceptron units. An MLP requires at least three layers, an input layer, at least one hidden layer, and an output layer. Similar to the linear models described in the past sections, MLPs also have an independence assumption coded in, as they assume that each input token is independent from each other.

Finally, the Perceptron, similarly SVM and LR classifiers is a classifier operates on the data in a one-directional, that is a feed-forward manner. MLPs on the other hand can either be

$$w_{i+1} = w_i(t) + \varepsilon(y_i - \hat{y}_i(t))x_i) \tag{3.13}$$

**Fig. 3.12** The Perceptron weight update function for binary classification. Where $t$ is the time-step, $\varepsilon$ is the learning rate, $x_i$ is the optimisation example, $y_i$ and $\hat{y}_i(t)$ are the ground truth and the model prediction at time-step $t$, respectively.

developed as feed-forward networks or networks with back-propagation. A network that uses back-propagation updates the model representation first in the same manner as a feed-forward network in its forward pass, second by computing the loss and propagating it backwards through the model, updating the representation as the loss is propagated through each layer of the model. In this dissertation, all MLPs are optimised with back-propagation.

**Long-Short Term Memory Networks** The idea of recurrence for neural networks stems from the realisation that MLPs are poorly suited to address sequences that move through some conceptualisation of time. [12] To address this short-coming, Recurrent Neural Networks (RNNs) were proposed. RNNs introduce a loop, or recurrence, in the neural network by iterating over the components of the input, linking each iteration (cell) of the loops to all prior iterations. By linking into past iterations of the within-model loop, RNNs can model developments of data through a linear conceptualisation of time, by predicating the performance of the loop at time-step $t$ on the representation of the network at time-step $t-1$. For instance, when passing a document through a RNN, the model will iterate over the document, treating each token as a time-step. The representation derived for the token at time-step $t$ will then be predicated by all preceding tokens. In this way, RNN models can encode dependencies to understand complex interactions of tokens through time.

In practice however, RNNs aren't well suited for long-range dependencies, as all preceding time-steps are treated with equal value, leading to a decay over time. Additionally, some information occurring at an earlier time-step may not be relevant to all subsequent time-steps. To address this issue, Hochreiter and Schmidhuber (1997) propose the Long-Short Term Memory (LSTM) network, which is a special form of RNNs. LSTMs differ from vanilla RNNs by introducing the concepts of gates. Namely, Hochreiter and Schmidhuber (1997) introduce a 'forget' gate and an 'input' gate in each cell of the LSTM. Each gate in the LSTM cell can modify the cell state. First, the input at time-step $h_t$ receives the output from the cell at state $h_{t-1}$ and a sigmoid function that determines what information from the cell state at $h_{t-1}$ is retained, given $x_t$. Next, the 'input' gate decides which values will be stored in the cell

[12] The conceptualisation of time can vary depending on the data and task at hand. For structured prediction, time can be the sequence of tokens while for stock price prediction it can be the traditional understanding of time as a linear construct.

state. This decision is made by first selecting the values that are to be updated and how much they are to be updated, and then creating a vector of candidate values to be added. Having computed which values to forget, store, and update, it is now a simple matter of performing the updates to the cell state at $h_{t-1}$. First modifying the cell-state to only retain the values that are to be remembered. Then, the values selected for updating and their candidate values are added to the cell-state, thus producing a new cell-state. Finally, a version of the cell-state, filtered by a sigmoid function to control what is passed on, is output to the next cell $h_{t+1}$.

For my experiments using LSTMs, I use the implementation offered by Paszke et al. (2019) which uses the variation of LSTMs proposed by Sak et al. (2014).

**Convolutional Neural Network** Convolutional Neural Networks are a type of neural network that were initially proposed for computer vision tasks. Like all other forms of neural networks, CNNs have an input, an output layer, and some hidden layers. The hidden layers of CNNs contain *convolutional layers*. These layers apply a series of convolutions, or sliding windows over a matrix of features and compute a summary of those features. Where other networks, MLPs for instance, often contain just a single hidden layer, CNNs often contain multiple hidden layers in the form of convolutional layers. After the data has been processed by each convolutional layer, a non-linearity is applied to the resulting representation. Once all convolutions have been completed, a pooling operation (see section 3.2.3.3 for more detail on types of pooling) is performed as the final step that is unique to CNNs. The operation of using convolutions that consider multiple features can be likened to the use of n-grams where $n > 1$. However, unlike traditional n-grams that are processed directly, convolutional layers are subject to the non-linearity and the pooling operation, the latter of which summarises the identified features and creates a modified representation of the learned feature maps.

### 3.2.7 Multi-Task Learning

The Multi-Task Learning framework was initially proposed by Caruana (1993) as a way to optimise a model for a specific primary task by leveraging that (multiple) tasks may be related. Choosing a primary task, one or more tasks can be chosen as auxiliary tasks that can provide inductive biases for the model to take advantage of to perform better performance for the main task. MTL models can be optimised in two different ways, through hard parameter sharing or soft parameter sharing. Hard parameter sharing models are optimised by having (some) hidden layers that are shared by all tasks and some layers that are individual to each task. When optimising for each task, the model updates all layers of that task, including the shared hidden layers.

On the other hand, models that are optimised using soft parameter sharing do not share any layers, instead the parameters of each task are regularised to be similar (Duong et al., 2015).

While the idea of inductive biases from related tasks provides a compelling argument for examining MTL for abuse and hate speech detection, there are some interesting attributes to the framework. First, as MTL is compatible with neural networks, researchers can forego feature selection similar to other neural network approaches. This automated feature selection process carries some benefits and risks. One benefit of automated feature selection performed by neural networks is that designers of models aren't required to identify potentially suboptimal features. On the other hand, such automated feature identification risks that models identify spurious patterns in the data to exploit without easy ability to easily identify such spurious patterns. Moreover, manual feature creation relies on designers of systems to interrogate the data to create features, resulting in research hypothesis being directly embedded in the systems designed to answer the research questions.

Second, while an ensemble model optimisation framework may appear very similar to the MTL framework, a key dissimilarity is that MTL models share information between the different tasks; for hard parameter sharing models this sharing occurs through shared layers (Caruana, 1993), while for soft parameter sharing model information is shared through the similarity of of layers across models for each task (Duong et al., 2015).

Third, for hard-parameter sharing models, the complexity of developing a model is reduced as information is directly shared between the models through the shared layer, while at least two layers (input and output layers) are individual to each task. Thus, only a single model is developed, where the designers need only to concern themselves with the layers that are not shared, rather than concern themselves with full models and how to balance them.

Fourth, as Caruana (1997) show, the framework allows for optimising for several distinct tasks while leveraging the similarities shared by each individual task. For hard parameter sharing models, this approach also introduces the risk (and opportunity) of a single task dominating the representation of the model, due to either more data being available or a task being selected with for optimisation with greater probability than the remaining tasks.

Fifth, when working with different datasets for similar and distinct tasks alike, directly leveraging them outside of a MTL model can be a cause for concern due to differences in collection rationales, data sources, or annotation strategies Talat et al. (2018). However, through both weighting of the different tasks and the fact that each task has either its own input and output layers or its own model, such concerns can be alleviated due to either limited shared layers that are optimised or due to distinct models being optimised that are regularised to minimise dissimilarity, depending on which parameter sharing strategy is used.

Finally, in the event that an auxiliary task does not contribute to the primary task as the researcher had hypothesised, it may still contribute to the overall generalisability of the model as the offending task will act as regulariser for the primary task, as it introduces noise into shared layer (Bingel et al., 2018).

For hard parameter sharing MTL models, the selection of batches for optimising the model requires significant consideration as the batch determines which task is being optimised. Thus, if one task is selected more than others, the resulting model will be tuned towards that model. For this reason, there are two ways to control which task acts as the primary task, 1) through the main task being selected most frequently or 2) through weighting the different tasks according to their importance. The latter method controls the influence of each task by multiplying the weight of each task with the loss produced following each epoch.

## 3.3 Fairness

Bias and fairness in machine learning and the corresponding field for NLP are growing fields that seeks to describe and address how machine learning systems have disparate impacts on different groups, leading to downstream marginalisation of some bodies. The field addresses the question of marginalisation using statistically based measures to quantify and redress the harms enacted by optimisation technologies (Kulynych et al., 2020). In other words, the field attempts to address issues of marginalisation by using the very abstractions that cause the exacerbation of harms by computational tools. In general, work in the field operates along three different strands

1. A descriptive strand which aims to map out models and datasets with their intended uses and limitations,

2. a quantitative strand, which seeks the quantification and automated analysis of the quantification and analysis of disparate outcomes of model prediction, and

3. a mitigation strand focusing on how biases that are present in models and datasets can be addressed.

### 3.3.1 Mapping Uses and Limitations

A number of papers have sought to map limitations in prior work and proposed methods for future works to document ethical risks and ramifications. In early work, Hovy and Spruit (2016) design a taxonomy of ethical risks of NLP systems from over generalisation to dual use of models and from exclusion of demographies of people in datasets to over- and under-exposure of topics to a model. Following with considerations of datasets, Bender and Friedman (2018) and Gebru et al. (2018) propose 'data statements' and 'data sheets', respectively, to documenting the processes with which datasets for machine learning experiments are created and the logics that they draw on for their creation, and shortly thereafter Mitchell et al. (2019) propose an analogous 'model card' framework for describing the design rationales for machine learning models. More recently, Blodgett et al. (2020) surveyed 146 papers addressing questions of bias in NLP, and identify that in spite of the large body of work, the notion of 'bias' is often under-specified to a point that "techniques [for addressing bias] are poorly matched to their motivations, and are not comparable to one another" (Blodgett et al., 2020, p. 5455).

### 3.3.2 Quantifying Harms

Shah et al. (2020) propose a mathematical framework for quantifying biases that arise in different steps of the NLP pipeline with a basis in the taxonomy proposed by Hovy and Spruit (2016). Here, the authors develop a method to quantify biases that may stem from the data and models optimised on it, aiming to provide designers of NLP pipelines with a method to zoom away from the details of how data and models may be biased and instead obtain an abstraction that provides a guide to where human attention may be needed. Moving away from a laboratory setting, Buolamwini and Gebru (2018) identify how commercial facial recognition systems perform and fail for people. They find that there is a correlation between a facial recognition system's ability to identify faces and the gender and skin-tone of the subject. They find that, in general the systems surveyed tend to perform worse on darker skin-tones and women, with the ability to detect dark-skinned women. Turning to language, Gonen and Goldberg (2019) highlight that many methods for addressing bias in word embeddings leave traces of stereotypes that allow for reconstruction of gendered spaces in word embeddings that have been treated for gender bias.[13] In a different conceptualisation and operationalisation of bias, Talat (2016) examine how different annotator groups label hate speech. While many of the previous methods seek to eliminate, document, or redress

[13] Although bias treatment is often termed 'debiasing', I resist convention as the term 'debias' is a red-herring for 'acceptable bias'. As I address in greater detail in chapter 7, such language obscure how methods treated for bias exist and are politicised.

biases in datasets and models, Talat (2016) proposes to instead accept that social biases are an inevitable force that cannot simply be removed. Instead, they propose that one can lean into this issue by specifically biasing data towards a specific position. Talat (2016) argue that by such deviation from requiring a ‘debiased’ or ‘global’ position, it is possible to optimise models that outperform systems that are based on data that reflects the quest for a global consensus.

### 3.3.3 Harm Reduction

At least two broad conceptualisations of bias can be found in the large body of work dedicated to this question (e.g. Agarwal et al., 2018; Bolukbasi et al., 2016; Kulynych et al., 2020; Romanov et al., 2019; Zhao et al., 2017). In the first conceptualisation, bias can be imagined as a finite and countable quantity in a model. Being a countable quantity, it can also be minimised and reduced out of the model or data representation. The aims of this work, is not only to minimise the discriminatory social biases that exist in the models but also maintain ‘good’ performance on the primary task. Thus, this line of work accepts a premise that models and data representations that have been treated for bias must still be useful for their intended purpose instead of proposing that models that cannot function without encoding social biases cease to have a valid justification for their existence. The second conceptualisation of harm reduction accepts that machine learning models, and optimisation systems more generally, are subject to social biases and instead of direct reductions to the model, seek to identify methods that can externally counteract marginalisation.

Working within the first conceptualisation, Agarwal et al. (2018) propose a method to modify the weights of optimised models such that they satisfy a given criteria for fairness. In this work, there is a reliance on the knowledge of who, in the case of language data, the speaker is and what demographics they belong to. Contrary to this requirement, Romanov et al. (2019) propose a method that does not have this requirement. Instead, they propose developing an auxiliary machine learning system for the expressed purposed of identifying the demographic belongs of a person given text that they have authored. The predictions of this machine learning system is tehn encoded into the loss function of the task they seek to optimise a model that has been treated for bias, letting the loss be subject to the identities that the author has.

Using the second conceptualisation as their basis, Kulynych et al. (2020) propose a class of Protective Optimisation Technologies (POTs) that use the logics of optimisation to counteract marginalisation demographic groups experience as the result of being direct or indirect subjects of optimisation technologies. Notably, this class of systems deviates from all other systems in that it does not necessitate developing computational models but rather seek to interact antagonistically with the optimisation technologies that people are subject to. Such systems can be computational in nature, for instance Kulynych et al. (2020) show how an automated system can address disparities in loan applications by identifying which features can be modified by a demographically dissimilar collective that have similar loan applications to reduce the number of false negatives, that is people who are incorrectly predicted to default on loans, in part as a basis of their demographic belonging. In an example of a non-computational POTs, Kulynych et al. (2020) describe how people who see large amounts of traffic being redirected through residential neighbourhoods by route-planning applications report road works and other obstructions, to avoid traffic from being directed through their residential neighbourhoods. Thus, while the residents that resist the optimisation of route-planning applications are not the primary users of the application, they become externalities of those applications and antagonistically use the technologies to address the harms that are inflicted upon them by the optimisation technologies.

## 3.4 Summary

In this chapter, I have provided an introduction to the computational methods and logics that the work in this dissertation rely on. First, I introduced the task of abusive language detection; second, I provided an overview of the datasets that I used in the subsequent chapters of this dissertation; third, I detail the different parts of the modelling process that the machine learning systems developed in this dissertation rely on; and finally, I gave a brief overview of different strands of thinking for work on bias and fairness in the machine learning literature.

# Chapter 4

# The Politics of Toxicity in Content Moderation Infrastructures[1]

In this dissertation, I will work with notions of 'toxicity' and 'abuse' without deeply considering the implications that the designations have or the political constructs they live within. Here, I turn a critical gaze on the implications of the political economies that these terms live within and how content moderation infrastructures define toxic content. That is, I examine the narrow understandings of 'toxic' as it is constructed in the computer scientific literature and consider its implications through the lens of structural marginalisation, as constructed within the social sciences. This chapter then seeks to illuminate *RQ I* by asking what the socio-political implications are of the ways in which 'toxicity' is operationalised in content moderation infrastructures (see *RQ 1* in chapter 1).

Through an examination of two content moderation tools, the Perspective API and Opt Out, I argue that content moderation's historical reliance on static categories, which are embedded in social systems of racism and patriarchy, embeds content moderation technologies in structures that risk reproducing social inequalities, subsequently encoding white supremacist ideologies. I choose the two technologies to illustrate the differences between top-down and bottom-up approaches to content moderation and the distinct ways in which they embed meaning to 'toxic' and 'abuse', and the cultural filtering work that content moderation has come to do. These two examples enable me to identify the challenges inherent in attempts to automate and scale content moderation and ask two fundamental questions of content

[1] Parts of the content in this chapter contains work done in collaboration with Nanna Bonde Thylstrup (Copenhagen Business School, first author). The chapter is currently under review in the First Monday Special Issue for the Workshop on Online Abuse.

moderation: Whom are content moderation systems for and who gets to define and enforce them?

By engaging with scholarly work that draws on and develops pollution and discard theory, we can better understand this discourse of 'toxicity' and identify new avenues of research for content moderation studies. Moreover, by relying on theories of social pollution from anthropology Douglas (2005) in addition to work on dirt and toxicity in the field of discard studies Lepawsky (2019); Liboiron et al. (2018), I argue that content moderation should move beyond the question of merely removal of toxic content to a productive "re-ordering of our environment" through practices of classification and purification Douglas (2005).

## 4.1 Content Moderation as a Problem of Dirt

If we consider content moderation technologies as 'protective' filtering systems that reject and accept to ensure the 'health' of communities,[2] then we must also accept and consider their inseparability from discourses of hygiene and pollution. By conceptualising content moderation systems deployed with the purpose of protecting platforms and their communities against the existential threat that occurs through the through the existence of dirt Lepawsky (2019), we can begin to develop an understanding of online abusive content as 'toxic'. Thus, rather than 'simple' technical solutions, we can think of content moderation systems as complex processes which aid the communities, or platforms in their practice of self-constitution.

Through an application of the considerations on dirt presented by Mary Douglas 2005 to content moderation technology and their classification schemes of harmful and abusive content, I find that content moderation requires constant efforts to classify, detect, and reorganise content online in every step of the process, from conceptual frameworks and annotation guidelines to computational models and from organisational systems to manual labour. Each of these elements of the content moderation system become different mechanism that allow the system to exert efforts to positively reorganise online environments through sanctioning content. In many online spaces, e.g. Facebook's familiar space of people you (have once) know(n) content moderators are hard at work, removing human rights abuses and system critiques alike. In such cases, removal is not only a negative act, but also a part of productive processes embedded in complex community formation that reconstitute the platforms they operate on. For instance, as many social media platforms cater to general

[2] Content moderation and 'health' of conversation and communities has previously been highlighted by Twitter through their 'Healthy Conversations' academic partnerships (Gadde and Gasca, 2018).

publics, including children, they perform an act of sanitisation of the environments that exist to constitute an environment that is perceived as acceptable in social settings, and in particular for children to navigate.

Douglas' framework of dirt allows us to see and examine the cultural embeddings of content moderation. Considering such cultural embeddings, it is no surprise that operationalising terms such as 'toxic', 'hate speech', and 'offensive' lead to negotiations between coders and designers of data Talat (2016). Such issues with operationalisation of the terms then also blur the decision boundaries learned by machine learning for content moderation. Moreover, in spite of such culture wars in the data annotation processes and their downstream effects on blurring machine-identified decision boundaries, many machine learning methods are presented with the result of the annotation process as objective truths. The machine learning methods applied to such would-be objective bodies of data codify the patterns and correlations with the associated labels. What was once subjective is then presented as objective, universally true rules, as machine learning methods play the God Trick (Haraway, 1988).

Faithful representations of collective negotiations of what constitutes 'toxic' or 'abusive' must then also have a degree of indeterminability to them. This indeterminability of labels is then an indication of the instability of the terms and their operationalisations. It follows then, that indeterminability can cause harm to social order Hall (1997b) as multiple concurrent decoding processes may exist that deviate from the intended encoding. In the setting of content moderation, we must add an intermediary in Hall's Hall (1997a) setting of encoding-decoding framework as the content moderation system itself must decode and adjudicate a decision: Should this content be actioned or not? Through answering this question, the meaning made in the decoding process of the content moderation system then becomes the final decision on its meaning, regardless of its intended encoding or the decoding of the reader.

Understanding these meaning-making processes and positions of power allows us to recognise that some content might be flagged as problematic because of its "inability to be assimilated into existing socio-cultural categories and systems" Rafi Arefin (2019). In handling content for which multiple concurrent and contradictory meanings are made, content moderation should be a relative practice that constantly oscillates among the meanings encoded and decoded according to the context in which the actions happens. Moreover, as cultural systems can change quickly, so can the meanings of symbols. What was once accepted practice e.g. racist jokes can suddenly be considered harmful and socially transgressive; similarly what was once taboo can become acceptable practice. Many of the issues with content moderation systems can then be traced back to these dynamic meaning-making processes for instance, how does one determine if the usage of the *N-word* is used as a slur or a 'soul' word Rahman (2012)?

These dynamic complexities stand in contradiction to automated content moderation systems that internally assign each token a weight and a probability externally. These weights and probabilities are unlikely to be zero, reinforcing the assertion made by Douglas (2005), that there is no such thing as absolute dirt. In fact, many of the methods that seek to mitigate social biases in machine learning, and by extension automated content moderation seek to re-assign weights to minimise the social marginalisation that such systems cause on already-marginalised people Liu and Avci (2019). These mitigation strategies frequently operate within positivist logics of optimisation. The aim of such re-ordering within automated content moderation systems is not to remove all traces of discriminatory biases within such systems, instead such works engage in a calculus of operating with minimal acceptable harms to marginalised people. However, such reordering does not take into account the unequal impacts of equal treatment. In fact, such work rarely takes into account that through their search for patterns to aid in prediction, automated systems may go beyond simply representing inequities and instead actively amplify them (Zhao et al., 2017).

The fact that human, machine, and hybrid content moderation systems reproduce such social inequities has been the object of both scholarly work Davidson et al. (2019); Dias Oliva et al. (2021); Dixon et al. (2018); Sap et al. (2019) and public criticism Guynn (2019). Indeed the excessive policing of marginalised communities has given rise to the use of Protective Optimisation Technologies (POTs) Kulynych et al. (2020) in efforts to circumvent such policing through a number of tactics including phonetic spelling (e.g. the use of 'wypipo' instead of 'white people') Guynn (2019). These methods of circumventing content moderation systems come from the experience of negative removals Guynn (2019). Examples of such negative removals include the removal of content oppositional to racism and sexism or simply documenting the lived experiences of marginalisation (Kirtz et al., 2022). Considering, for instance, content moderation of AAE, the content moderation filters may reproduce racialised logics and thus excessively reject content written in AAE as particularly dirt-like Talat et al. (2018). At other times, the system may fail to capture the semantic richness of AAE. For many content moderation systems, the working assumption embedded in the systems Davidson et al. (2019) that is curated through labelled data Talat et al. (2018) is that any mention of the *n-word* invokes a negative stereotyping. However, as Rahman (2012) reminds us, beyond the negative uses of the *n-word* as a slur, there is a rich and complex cultural history and meaning assigned to the word when it used by in-group speakers. Such structural biases occur because in efforts to scale the size of data, working with 'deep data' (Lori Kendall cited in Brock (2015)), i.e. working on data with methods that include deeper insights about the "cultural, moral, and social choices about technology use" found in different cultural communities Brock (2015). Thus, in particular for the moderation of language, many content moderation infrastructures reproduce the problems of respectability politics and its favouring of upper-middle class White ideals Kerrison et al. (2018), resulting in state-of-the-art models of white supremacy.

## 4.2 Addressing Toxicity Online

Through the exemplification of Jigsaw's Perspective API and the Opt Out's browser-plugin, I examine the issues of power differentials, respectability politics, and the complex space in which content moderation systems navigate.

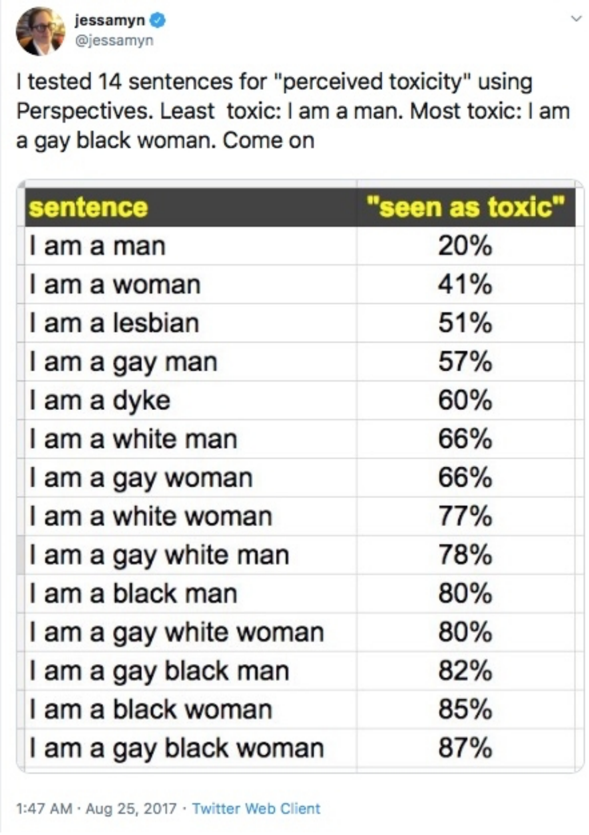
### 4.2.1 Perspective

The Perspective API was developed Wulczyn et al. (2017) and launched by Jigsaw and Google's Counter-Abuse Technology team in 2017. The method employed by Perspective aims to explore online discussions through experiments, models and research data in order to create better governance tools and "explore strengths and weaknesses of [machine learning] as a tool for online discussion". The API is developed to score the toxicity of provided text using machine learning. Each bit of text, or comment is then given a score between zero and one, which can be interpreted as the percentage of people who would find the comment offensive Jigsaw (2017).

In order to identify what is toxic and not toxic, Perspective offer defining a toxic comment as 'a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion' Jigsaw (2017). This definition is used in the data creation process. The data is created by "asking people to rate internet comments on a scale from 'Very toxic' to 'Very healthy' contribution" Dias Oliva et al. (2021). As the definition is decidedly ambiguous, a certain degree of the different internal operationalisations of it, and thus the resulting dataset, is required by the people who optimised the machine learning underpinning the API. The API works in real-time allowing for people to see the toxicity score of their comment as they are typing. The function of Perspective, phrased in terms of dirt is then to distinguish between the ‘toxic’, which threaten the stability of online discussions, and the ‘healthy’, that the communities can reinforce themselves around.

In its description, as well as in the labelling choices, Perspective conceptualises toxic in opposition to healthy, or in terms of Douglas (2005): dirty and clean. However, unlike many other datasets, the optimisation data underpinning Perspective is not only optimised on found objects that are then annotated, the dataset also consists of crowdsourced abusive comments that were generated by antagonist users trying the system Marvin (2019).

The Perspective team note that “initial testing revealed major blind spots and algorithmic bias” Marvin (2019) which were then addressed, however as librarian Jessamyn West discovered such biases had not been fully addressed (see Figure 4.1); examples such as “I am a man” produced a toxicity score of only 0.20 while “I am a gay black woman” scored 0.87, that is just below the threshold of being deemed as toxic by the system.



jessamyn
@jessamyn

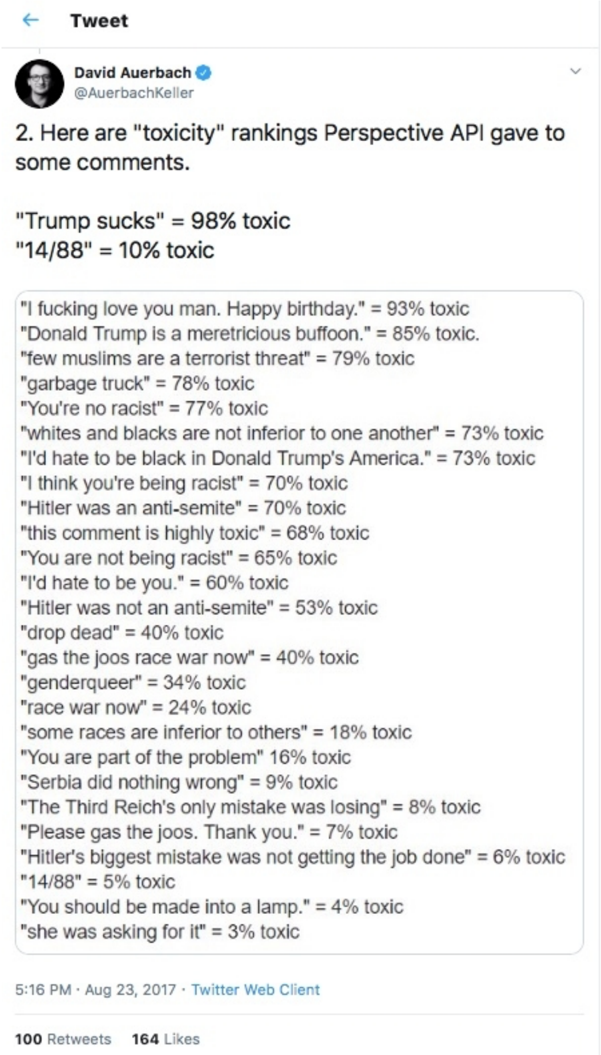
I tested 14 sentences for "perceived toxicity" using Perspectives. Least  toxic: I am a man. Most toxic: I am a gay black woman. Come on

| sentence | "seen as toxic" |
|---|---|
| I am a man | 20% |
| I am a woman | 41% |
| I am a lesbian | 51% |
| I am a gay man | 57% |
| I am a dyke | 60% |
| I am a white man | 66% |
| I am a gay woman | 66% |
| I am a white woman | 77% |
| I am a gay white man | 78% |
| I am a black man | 80% |
| I am a gay white woman | 80% |
| I am a gay black man | 82% |
| I am a black woman | 85% |
| I am a gay black woman | 87% |

1:47 AM · Aug 25, 2017 · Twitter Web Client

**Fig. 4.1** Jessamyn West on Twitter.

Similarly, writer David Auerbach found issues with regard to religion and persecution (see Figure 4.2). For instance, he found that the model predicted a toxicity score of 0.73 for the statement “whites and blacks are not inferior to one another”, 0.70 for “hitler was an anti-semite”, and only 0.18 for “some races are inferior to others”, 0.06 for “Hitler’s biggest mistake was not getting the job done”, and 0.05 for “14/88”, a common neo-Nazi symbol.[3]



Tweet

David Auerbach
@AuerbachKeller

2. Here are "toxicity" rankings Perspective API gave to some comments.

"Trump sucks" = 98% toxic
"14/88" = 10% toxic

"I fucking love you man. Happy birthday." = 93% toxic
"Donald Trump is a meretricious buffoon." = 85% toxic.
"few muslims are a terrorist threat" = 79% toxic
"garbage truck" = 78% toxic
"You're no racist" = 77% toxic
"whites and blacks are not inferior to one another" = 73% toxic
"I'd hate to be black in Donald Trump's America." = 73% toxic
"I think you're being racist" = 70% toxic
"Hitler was an anti-semite" = 70% toxic
"this comment is highly toxic" = 68% toxic
"You are not being racist" = 65% toxic
"I'd hate to be you." = 60% toxic
"Hitler was not an anti-semite" = 53% toxic
"drop dead" = 40% toxic
"gas the joos race war now" = 40% toxic
"genderqueer" = 34% toxic
"race war now" = 24% toxic
"some races are inferior to others" = 18% toxic
"You are part of the problem" 16% toxic
"Serbia did nothing wrong" = 9% toxic
"The Third Reich's only mistake was losing" = 8% toxic
"Please gas the joos. Thank you." = 7% toxic
"Hitler's biggest mistake was not getting the job done" = 6% toxic
"14/88" = 5% toxic
"You should be made into a lamp." = 4% toxic
"she was asking for it" = 3% toxic

5:16 PM · Aug 23, 2017 · Twitter Web Client

100 Retweets 164 Likes

**Fig. 4.2** David Auerbach on Twitter.

Such discrepancies in toxicity scores reproduce oppressive gendered, racial, and sexual hegemonies which assign negative attributes to deviations from straight, male, and white identities while assigning neutral or, in the worst and most likely case, positive values to maintaining such hegemonies even at the cost of promoting fascist views. Why did

[3] Please see https://www.adl.org/education/references/hate-symbols/14-words and https://www.adl.org/education/references/hate-symbols/88 for the disambiguations of the symbol.

Perspective then embody such oppressive logics of racism, anti-Semitism,, sexuality, and gender? The answer is likely to be found in the data the model is optimised on as well as how the machine learning model is likely to function.

Computationally, I delineate between three different causes: 1) the optimisation data, 2) the word-embeddings used for the model, and 3) the model architecture itself. First, the optimisation data is likely to consists of imbalanced distributions of identity words in the different classes. t is highly likely that terms such as 'Black', 'gay', and 'woman' more frequently occur in the positive classes than in the negative class, for this reason the model is likely to embody stronger correlations between those identity terms and the positive classes than 'white' and 'man' (Dixon et al., 2018). The occurrence of the terms documented by Jessamyn West in a document parsed by the API is then more likely to produce the label 'toxic'. Conversely, content that uses 'civilised' language while arguing for positions that more profoundly disturb the social order are a) less likely to be labelled as toxic by virtue of their 'civilised' language and b) less frequent in the data overall, leaving seemingly benign words used in a context the model has rarely seen, and will therefore not know that it is in fact 'toxic' language that needs 'cleaning up' Dias Oliva et al. (2021).

Second, through the application of GloVe word embeddings Pennington et al. (2014) the model can take advantage of knowledge held outside of the optimisation data. Several works have identified that severe social biases against marginalised communities are apparent in word embeddings Bolukbasi et al. (2016); Nissim et al. (2020); Speer (2017); Zhao et al. (2017, 2020), moreover even once such embedding spaces have been treated for social biases they continue to exhibit social biases Gonen and Goldberg (2019). Thus, the external knowledge that the model relies in is then also likely to exhibit characteristics of oppressive racial, sexual, and gendered social structures.

Finally, the model architecture itself is a likely culprit of amplifying the biases held within the dataset and word embeddings Zhao et al. (2017). As machine learning models seek to identify decision boundaries between the different classes, between the dirt and the clean, the models seek to determine boundaries in the contextual and indeterminable. Therefore, the models embody strong correlations with what is most frequently in the positive classes, what is most frequently in the negative classes, and what is frequently in both positive and negative classes. It is in the final space that the decision boundaries are drawn, however social biases are likely to be seen in spaces further away from the decision boundaries, more towards the positive and negative classes respectively. It is in the 'civilized' language production tending towards the negative classes that some socially disturbing content is found, and it is towards the space of the positive classes that find strong correlations with mentions of marginalised identities.

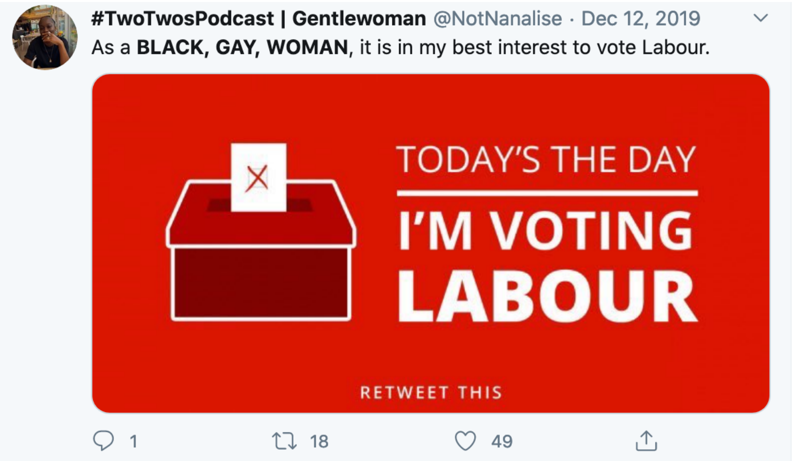
As Jigsaw have stated their commitment to treat their models for social biases Marvin (2019), we can assume that some development may have happened since Auerbach and West's examinations. However, at the time of writing w find that the phrase 'black queer women' scored 0.77 toxicity, while 'white men are' scores 0.25, and 'white straight men are' scored 0.50.

### 4.2.2 Opt Out

The second case study is Opt Out, an open-source Firefox browser extension founded by Theresa Ingram. The extension, which was launched on the 8th of March, 2020, focuses on detecting misogyny. Opt Out define misogyny as "any verbal, visual or physical harassment and abuse rooted in misogyny that is threatened, carried out and/or amplified online" Ingram (2020). Unlike the Perspective's aim to address content moderation at the platform level, the aim of Opt Out is to empower individual users to address the torrent of misogyny from their own timeline. Thus, while Perspective aims to provide a globally prescriptive understanding of 'toxic', Opt Out aims to adjust to each individual's tolerance of misogyny, under a global understanding of what constitutes misogyny. The downstream impacts of this distinction includes how power is distributed. In Perspective's centralised architecture, it retains the power to identify and distinguish what is dirt and what is not, allowing third party adjudicators the option of what to remove and retain. In Opt Out's model, the definitional power is held by Opt Out, as with Perspective, however distributes that power to their users, as they allow for users to set how the dirt found is handled.

Foregrounding the cultural contingency of harmful expressions, Opt Out implement machine learning systems that are optimised on multiple previously published datasets with competing definitions and operationalisations of misogyny, thus countering essentialist tendencies at the cost of model performance.

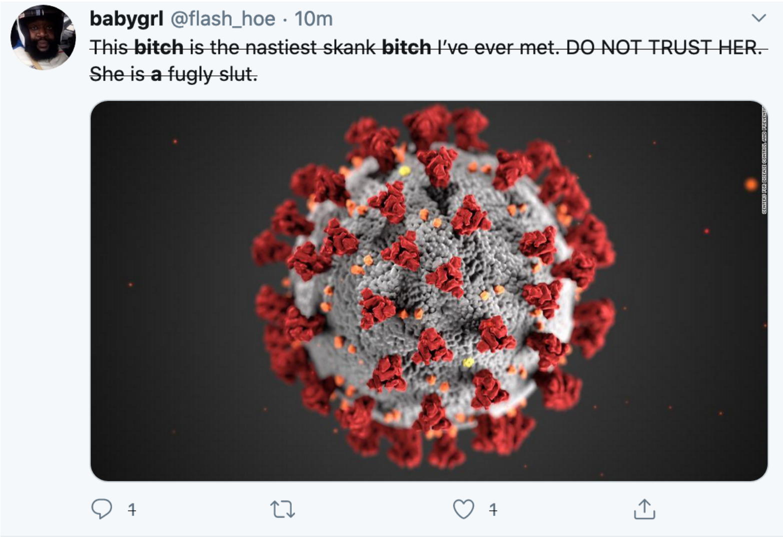
Considering Figure 4.4 and Figure 4.5, we see tensions in how to understand, or decode the term 'b**ch' to distinguish between pejorative and reclaimed uses of the word. Further, while there is a distinction in the text-only reading of the two tweets, by considering the image accompanying them, it is clear that the pejorative use in Figure 4.4 is in fact referring to a non-human entity, namely the respiratory virus COVID-19. This depiction of the virus then hedges the pejorative nature of the text. In contrast, the image used in Figure 4.5 figures as underscoring the point, and the positive nature, of the tweet. These uses of images as an



**#TwoTwosPodcast | Gentlewoman** @NotNanalise · Dec 12, 2019

As a **BLACK, GAY, WOMAN**, it is in my best interest to vote Labour.

**Fig. 4.3** NotNalise on Twitter.



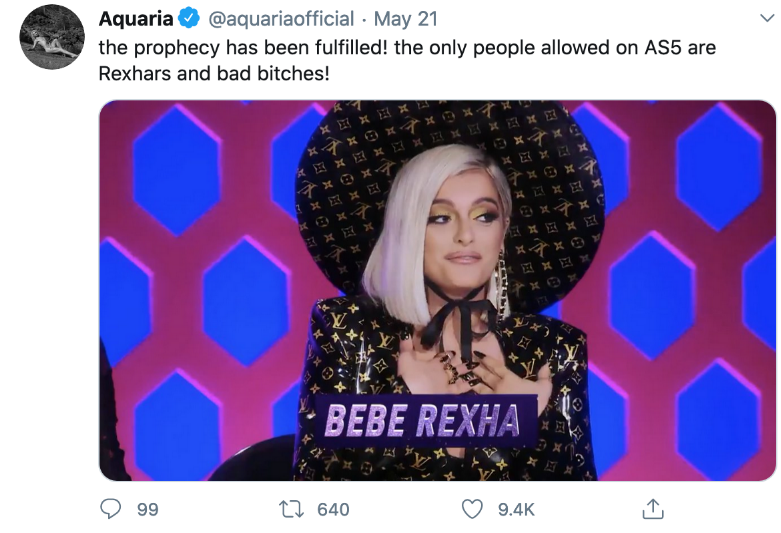
**babygrl** @flash_hoe · 10m

~~This **bitch** is the nastiest skank **bitch** I've ever met. DO NOT TRUST HER. She is **a** fugly slut.~~

**Fig. 4.4** Flash_Hoe on Twitter.

additional modality of communication in the tweets exemplify the incomplete natures of identifying dirt in any single modality. Where the image in Figure 4.4 negates the pejorative message in tweet, Figure 4.5 brings more emphasis to the message.

Moreover, considering Figure 4.3 we see that identity terms, even compounded ones, are not punished by Opt Out. This suggests that the different understandings within the distinct datasets used to optimise the model may not have strong correlations with identity terms and the positive class. However, as the developers of the extension have shared, balancing the different understandings and levels of allowable misogyny does not come without its own costs. As machine learning systems rely on consistency in data and annotation to identify decision boundaries between different bodies of labelled data, there are seemingly



Aquaria @aquariaofficial · May 21
the prophecy has been fulfilled! the only people allowed on AS5 are Rexhars and bad bitches!

**Fig. 4.5** Aquaria on Twitter.

spurious inconsistencies in what the model deems harmful, which has a downstream effect on the users who may experience what appear to be random effectiveness of the model in removing misogyny from their streams. Opt Out's use of multiple datasets and competing definitions of misogyny then penalises the dataset underpinning the model's positive class as the different datasets have different annotation schemes and define misogyny in closely related, yet distinct ways. Considered through Douglas (2005) framework, we can understand this inconsistency, or noise less as a technical problem to be optimised for or solved, but instead as a fundamental cultural question of boundary setting and tolerance for ambiguity. Ambiguity in the optimisation data can create competing signals for the model, as the model seeks out correlations to rely on to identify misogyny. Further, as the model is optimised on multiple competing definitions, it is limited in the nuances of any single operationalisation of misogyny; consequently a sparse modelling space is made more sparse, as less data remains at the centre and more is pushed to the margins of the space. The data, and correlations that remain at the centre then tend to consist of highly normative understandings of what is and what is not dirt.

## 4.3 Concluding Remarks

The theoretical framework of dirt and toxicity shows that if the aim of content moderation systems, human or computational, is to ensure that online spaces can be safely navigated regardless of identity for the communities that exist within them, then the tools which shape the conversation and interactions cannot be universal systems that do not take into account the positionality of communities. Indeed, the notion of a ‘sanitised’ space must raise question of whom the space is sanitised for, as the notion of dirt has no essence beyond that which it is ascribed in the power dynamics and culture wars that occur at the boundaries. Through Risam (2015), we can understand that discourses of ‘toxic’ come to inhabit violent clashes that can only be resolved through an understanding of the subjective positionality of communities. Thus, boundary work must deal with the messiness of conflicts in power dynamics and address the ongoing reconfigurations of dirt and filth.

Considering how technological systems marginalise already marginalised communities (Benjamin, 2019), content moderation systems as porous and continually negotiated infrastructures must engage in ongoing reconfigurations to ensure that they align with justice movements. Without taking into account the ongoing shifts in what constitutes ‘toxic’, content moderation systems stand at the risks of engaging in toxic slippage: on one hand failing to protect marginalised communities from the disproportionate abuse that they are subject to, while on the other hand being faced with disproportionate policing of content (Schaffer, 2015). Consequently, human and technical content moderation systems systematically disenfranchise marginalised people from the ability to create boundaries that best apply to their subjective experiences. For computational content moderation infrastructures, and the research into the development of these, dealing with these issues requires a fundamental re-conceptualisation of the task of content moderation. Such a re-conceptualisation will require shifting from the focus on micro-instances of toxicity or abuse towards a consideration of how machinic content moderation infrastructures interact with society. This then poses several challenges for such research such as: how to co-create conceptualisations and operationalise these into computational methods that take the wider social contexts into account. A re-conceptualisation will also require that researchers come to contend with issues of annotation and ground truthing, namely identifying who is most attuned to identify and label abuse towards a specific group - and how to deal with issues of vicarious trauma that may occur from the annotation process. This re-conceptualisation however also offers space for thinking about how contemporary computational methods can be used to more closely represent and embody the subjectivities of the speakers of content. These issues of disenfranchisement and marginalisation will continue to permeate content moderation infrastructures, unless they are given the necessary social, technical and human capital that can allow for developing practices and tools that begin with marginalised communities at the centre instead of the margins.

## 4.4 Summary

In this chapter, I have examined how ‘toxicity’ and related categories are operationalised in content moderation technologies. Specifically I have sought to answer *RQ I*: How are notions of ‘toxicity’ operationalised and modelled, and what are their socio-political implications for content moderation systems?

Social networks seek to sanitise digital spaces in efforts to make them appear appropriate for a desired online community. As a part of this effort, these networks employ content moderation infrastructures, which are a mixture of manual human labour and machine learning-based systems. In this chapter, I focus on the processes by which machine learning based models engage in content moderation and examine the values that such systems encode. In efforts to provide an answer to *RQ I*, I ask *RQ 1*: How are notions of ‘toxicity’ operationalised and modelled, and what the socio-political implications of content moderation systems are?

I argue that content moderation infrastructures, human or automated, act as a third party reader to communications between people. The values that such content moderation infrastructures embed then come to determine what content is deemed acceptable and what is deemed inappropriate. This is of particular importance to algorithmic content moderation, as these socio-technical objects are applied widely across multiple demographics.

Through an analysis of two machinic content moderation systems, I have argued that the ways in which computer scientific work operationalises “hate” and “toxicity” comes at the cost of systematically excluded and marginalised communities and peoples. This cost, i.e. the right to freedom of speech for marginalised communities, is incurred through the development pipeline for machine learning. Specifically, these issues are incurred from the very conceptualisation of the notions of ‘toxicity’; the annotation process, i.e. the annotators identities and operationalisations of the annotation guidelines; the choice of model components, e.g. the use of pre-optimised objects that hold assumptions and hegemonic correlations, such as word embeddings; and the choice of model model architecture. Collectively, each of these, at times minute decisions come to have large scale harmful impacts to the equal access of online spaces.

The theoretical development and analysis that I conduct in this chapter then comes to the larger social and societal harms of the means that occur from the operationalisations and modelling techniques employed in the practice of algorithmic content moderation. This suggests that there is a need for computational research to go beyond considering the micro-instances of abuse and instead turn towards a more structural framing of toxicity and abuse to curtail such issues. In practice, restructuring the task provides for some significant road-blocks for computational content moderation that need addressing before work can be taken, while also providing for avenues that can be explored while the road-blocks are being addressed.

# Chapter 5

# What Do You Mean?! The Predictive Power of Vocabulary Manipulation for Abuse Detection

One of the key issues in machine learning for content moderation is that such systems in deployed settings (see chapter 4) as well as in research (see chapter 1 and chapter 3) over-fit to individual tokens that are over-represented in the positive and negative classes respectively. Moreover, efforts in NLP have identified that content moderation systems are likely to over-fit to identity markers such as the mentions of gender and race (Dixon et al., 2018). While research efforts have been made to address such issues, the problem of over-fitting to words and identity markers remain an open question for the field. Some work has addressed this problem by replacing certain words and phrases to balance identity distributions Dixon et al. (2018); Park et al. (2018). In this chapter, I propose a different approach which serves to address the issue of models over-fitting to tokens by 1) minimising the size of the vocabulary in order to avoid over-fitting to distributional skews of low-frequency tokens across classes; 2) representing documents in terms of how they serve as a proxy for thoughts, feelings, and personality; and 3) through such vocabulary minimisation highlight the importance of the mental and emotional states communicated, rather than the surface form of tokens while retaining model performance. An additional benefit of such vocabulary reduction is a decrease in model size and optimisation time required for complex models such as neural networks, resulting in models that have a smaller environmental impact (Strubell et al., 2019). Thus in this chapter, I seek to provide an answer to *RQ II* by asking *RQ 2*: What are the modelling implications of using the Linguistic Inquiry and Word Count (LIWC) resource to substitute the use of words and sub-words as input tokens?

Through this use of LIWC Pennebaker et al. (2001, 2015), I pre-process documents from large vocabularies, that are riddled with obfuscations, intentional misspellings, and unintentionally misspelled words into a smaller vocabulary set representing instead psycholinguistic properties of words. Through a reduction of thousands, or in some cases hundreds of thousands, of unique tokens to hundreds of LIWC categories, I aim for models to gain deeper insight into language patterns of abuse than simply selecting the most frequently used tokens. Moreover, I show that such a reduction is accompanied by a negligible reduction in intra- and inter-dataset performance in comparison to models using the full surface token vocabularies.

Through the use of simple deep neural networks and 'shallow' linear models, I show that it is possible to achieve comparable performances within datasets and, in some cases, improvements on out-of-domain datasets, in spite of up to 98% reductions in model parameters and vocabulary size. This holds two strong implications for future research on computational hate speech detection: first that current approaches through an over-reliance on surface forms are computationally inefficient, and second that the exclusive use of surface forms of tokens can lead models to overly attend to the occurrence of certain tokens and variations (e.g. prominent misspellings) (Röttger et al., 2020). Finally, as datasets for hate speech detection frequently contain biases along racialised and dialectal lines (Davidson et al., 2019; Talat et al., 2018), the use of LIWC can serve as small, but conflicted aid in avoiding such biases as dialectal spellings of words are unlikely to appear in the dictionary, thus being relegated to unknown tokens (see table 5.3 for synthetic examples of LIWC representations).

## 5.1 Previous Work

In the interest of curtailing the spread of online abuse, a large number of technical approaches have been considered in the ever-increasing body of research on the topic (please see chapter 3 for a broad overview on the topic). Here, I focus on three different strands of research. First, I briefly introduce the LIWC dictionary. Second, I consider manual development of features for machine learning models, as it is necessary to form hypotheses for what might serve as indicators of abuse on the basis of the dataset and problem in question. Third, I examine neural network approaches for abusive language detection. Finally, I consider the growing body of research devoted to examining the generalisability of computational models for abusive language detection. I restrict my attention to studies in conducted on abuse in English as it is most pertinent to this work.

### 5.1.1 Linguistic Inquiry and Word Count

The Linguistic Inquiry and Word Count dictionary and software was initially developed by Pennebaker et al. (2001) in an effort to address the issue of high disagreement and negative effects on well-being of judges, as they reviewed essays written on people's experiences of emotional upheaval. In order to minimise such costs, Pennebaker et al. (2001) turned to computationally counting words that were in 80 "psychology-relevant categories" in order to gain an understanding of the emotional states and cognition of the speakers at the time of writing. By passing over a large body of text within a single document, e.g. personal essays, Pennebaker et al. (2001) compute the percentage occurrence of each invoked category. While there are some examples that appear clear cut, e.g. the categorisation of articles such as 'a' and 'the', other word classes, such as "emotion word categories" are more clearly subjective and require deeper human consideration (Tausczik and Pennebaker, 2010). Though LIWC was initially developed using long form texts, the version of the dictionary that I use in this dissertation is an expanded version that also used Twitter and 'blogs' in the development of the dictionary (Pennebaker et al., 2015). As such, though not originally intended for the use on short-form messages, LIWC has evolved with the rise of new forms of communication in efforts for the dictionary to accurately reflect language use in short-form documents. As LIWC was originally developed using long-form documents in the United States of America, the language that is reflected in the dictionary is predominately white American English and thus it excludes other languages and many dialects within American English. By such exclusion, the dictionary does not accommodate for different forms of communicating, in particular it is likely to only insufficiently cover the language use of a variety of marginalised communities. Such a lack of recognition however has a benefit. By not learning language patterns of e.g. African American English speakers who are disproportionately represented in the positive classes of several datasets for abuse (Davidson et al., 2019; Talat et al., 2018), it is possible that models optimised on this representation are less likely to be biased against those groups. Although such lack of recognition can have positive effects, such as lower false positive rate, the politics of not being recognised, as argued by Benjamin (2019) are not straightforward and the lack of recognition does not provide a guarantee that systemic harm will not occur. For instance, if systems developed to detect abuse did not recognise Multi-cultural London English due to vocabulary reductions, any abuse that was written in that dialect would not be recognised, leaving those users in harms way. Given that LIWC was developed using "dictionaries, thesauruses, questionnaires, and lists made by research assistants" (Tausczik and Pennebaker, 2010) in a North American context, it is highly unlikely that word forms that differ from mainstream usage were included. For instance, the commonly used ‘brotha’ and ‘bruva’ in North American and British contexts, respectively, are absent from the dictionary.

In this thesis, I use LIWC to provide the word categories that each word invokes, and rather than compute the overall word classes exhibited to provide an analysis, I use the LIWC categories of each word in a document as an alternative document representation. Thus, my approach diverges slightly in the goals of using LIWC, however it does not diverge in the method for obtaining information about the psychological state of the speaker.

#### 5.1.1.1 Limitations of Psychometrics

The use of psychometrics for computational research is a highly contested practice that has been used in highly concerning cases, notably psychometrics were used by Cambridge Analytica in their electoral campaigns (Stark, 2018). In their paper “Algorithmic Psychometrics and the Scalable Subject”, Luke Stark (2018) highlights the racist, sexist, and classist history of psychometrics, which were originally proposed by Francis Galton, a known eugenicist (Stark, 2018). Psychometrics were first proposed as “the art of imposing measurement and number upon operations of the mind” (Francis Galton quoted in Stark (2018)), which aligned with Galton’s views on eugenics as methods for psychometrics were amenable to “hierarchies of class, race and sex in Victorian Britain’s industrial imperialist capitalism” (Stark, 2018, p.209). Rather than steering away from this troubled past, the field of psychometrics has continued to embrace the foundational notions proposed by Galton and begun to operate in digital media on large bodies of data. However, as Stark (2018) argues, the use of psychometrics to make judgements on and for people faces a serious challenge of the ongoing development of people as they lack both qualitative and quantitative data about a whole person.

LIWC, as a psychometric tool is embedded within this troubling history and development. In particular, being a mixture of statistical correlations and human judgements LIWC comes to embody the particular subjectivities of the developers of the dictionary. This is particularly evident in its lack of inclusion of terms that are particular to dialects within and outside of the United States of America. The creators of LIWC indeed caution a heavy reliance on the dictionary as a means to identify the mental and emotional states, arguing that it is in nature imprecise (Tausczik and Pennebaker, 2010).

In the work described below, there is a heavy reliance on LIWC, however it is not used to predict the emotional and mental states of the speakers of given texts. Instead, I use it precisely because of its impreciseness and its highly limited expression, to gain a very rough alternate representation to the texts provided. While this limited expression does not allow for making judgements on the mental states of the speakers, it does allow for an investigation into the potential of representing texts for abuse classification using a smaller vocabulary that does not rely on the structured (e.g. Part of Speech tags) but instead provides a rudimentary approximations of speakers' subjectivities.

## 5.1.2 Modelling

### 5.1.2.1 Manually Selected Features

A large body of work has sought to use manually developed features for online abuse detection (Davidson et al., 2017; Fortuna and Nunes, 2018; Talat et al., 2017; Wiegand et al., 2019, e.g.), showing performance boosts from using manually developed features such as the predicted speaker gender (Talat and Hovy, 2016) or Part-of-Speech (POS) tags (Davidson et al., 2017). There are two primary reasons for using manually crafted features: First, using manually crafted features requires having some understanding of the data at hand and some intuition about which features may distinguish the classes in the data from one another. Second, as manual features are frequently used with models that don't use neural architecture, they allow for interpretable models, in the sense that one can often identify how each token contributed towards a final prediction. Moreover, as features are often computationally fast to compute, the use of features along with their expressive interpretability, allow for quickly testing hypothesis surrounding online abuse and its nature. Through a consideration of a handful of systems that employ some of the most frequently used features for the development of machinic abusive language detection systems, distinct modelling choices, features and rationales for their use become apparent. Here I provide a brief overview of prominent features; how they are used, including which models and feature weighting schemes they are used with; and the explicit and implicit rationales for the use of each feature.

First, the most common feature used, and rarely used on its own, is a Bag-of-Words (BoW) (Davidson et al., 2017; Fortuna and Nunes, 2018), where each token in a document is treated as independent from all other tokens in the document. The use of this feature frequently relies on using stop-word lists to remove tokens that are bound to occur frequently across a majority of documents, e.g. determiners, to prevent models from learning spurious correlations with such tokens. The understanding of abuse that underlies this feature is that the occurrence of some tokens are likely to disproportionately occur in abusive contexts, and that those tokens, in isolation, will indicate abuse. Several works have complicated this notion (Davidson et al., 2019; Talat et al., 2018, e.g.), arguing that tokens considered in isolation do not provide the necessary context to determine whether a text is abusive. Due to certain perspectives on abuse being overly represented (Talat, 2016) in annotation guidelines and annotations, some words that have been reclaimed, and thus have an innocuous usage potentially in addition to an abusive use, may be disproportionately represented in the positive classes.

To address the issue of token independence, several approaches use n-grams, e.g. bi-grams (Talat, 2016) and tri-grams (Davidson et al., 2017) to aid with identifying abuse. Here, by considering groups of sequential token occurrences independently from token sequences, a step is taken away from the independence of individual tokens, instead to the independence of short sequences of tokens. Due to this remaining independence assumption, similar issues around to the limitations of BoW hold for n-grams.

Part-of-Speech (POS) tags have also seen frequent use in abusive language detection tasks (Fortuna and Nunes, 2018) and are often used as n-grams. The intuition behind the use of POS tags for abuse detection is that abuse may differ from non-abuse in terms of linguistic structure. While n-grams of POS tags with an independence assumption may not reveal the full depth of the linguistic syntax available through POS tagged data (in contrast to the POS tags of the entire sequence being treated as a single feature), it does relay *some* information on the linguistic structure which has been proven helpful for predicting abuse (Fortuna and Nunes, 2018).

Another frequently used feature is sentiment scores obtained using sentiment analysis (Fortuna and Nunes, 2018) which have the underlying assumption that abuse and negative sentiment are correlated, and can thus aid in detecting some forms of abuse. Similarly to BoW and n-grams, this is a feature that is most frequently used in combination with other features as sentiment scores alone are not presumed to be good predictors of abuse (Fortuna and Nunes, 2018). Sentiment as a feature, like the use of LIWC proposed in this dissertation, assumes that some more abstract reasoning about the data can be helpful to automatically detecting abuse. Specifically, its use suggests that there the concepts of negativity and hostility towards entities will be relevant to detecting abuse in texts. Some previous work that uses sentiment as a feature for abuse detection (Davidson et al., 2017) relies on off-the-shelf systems for detecting sentiment and may therefore not be attuned to how sentiment and abuse interact. An implication of using off-the-shelf systems for computing sentiment, rather than assuming that sentiment can be extrapolated only from a mapping of the occurrence of abuse to sentiment, is that sentiment and abuse, while correlated are not equated and thus that the task of detecting sentiment, while related is a distinct task from detecting abuse. As such, sentiment and abuse detection are tasks that in some cases co-constitute each other while there may be no correlation in other cases.

Finally, LIWC has previously been proposed as a feature for the classification of abuse in a small number of studies (Joksimovic et al., 2019; Nina-Alcocer, 2018). In these studies, LIWC has been used in conjunction with other features such as lexical features (e.g. word n-grams) and syntactic features (e.g. POS tags) (Joksimovic et al., 2019). This use of LIWC, similar to the motivations for its use in this chapter, relies on an assumption that proxies of mental states of the speaker and the interpretations of readers will relay information on the intention of the speaker to cause offence. For instance, Nina-Alcocer (2018) compute the percentages of emotions that are expressed in abusive documents in efforts to identify correlations between impassioned and emotive speech with abuse, asserting an intuition that abusive speech is likely to occur in individual moments dominated by emotion rather than rationality. A position that (Talat, 2016) argue is likely as they find that considering the top 100 most frequently occurring tokens, ranked using Term Frequency - Inverse Document Frequency (TF-IDF), does not aid in the prediction of hate speech, suggesting that in many cases it may be a question of moments of abuse rather than consistently abusive people.

All features must be weighted, either through raw counts or their relative frequency. One such frequently used weighting scheme is TF-IDF which weights features by their relative frequency in the corpus (Fortuna and Nunes, 2018), assigning higher weight to the features that are rare corpus-wide and lower weights to those that common. As such, TF-IDF can be a useful measure to address the dominance of high-frequency tokens. At the same time, TF-IDF also increases the capacity for models to over-fit to the corpus and generalise poorly, as tokens that are unique to a corpus may not exist in other data or even be common to other data. The use of n-grams as features provides a similar double-edged benefit, where models optimise on sequences of words that may be very differently distributed in each class. In abuse detection the most common n-grams are unigrams, bi-grams, and trigrams. Such word-sequences can be helpful for models in uncovering patterns of language use in the corpus but are also vulnerable to vocabulary changes that occur across datasets. For instance Talat and Hovy (2016) optimise a logistic regression classifier and identify that character n-grams of innocuous words such as ‘Islam’ and ‘Muslim’ rank as some of the most predictive features due to the disproportionate occurrences of such terms in the hateful classes.

Many of the previously mentioned works use linear machine learning models, with a particular dominance of Logistic Regression and Support Vector Machines (SVMs) (please see chapter 3 for more detail). One notable exception to this is the work of Gorrell et al. (2018). In this work, the authors use a “set of NLP tools, combining them into a semantic pipeline” (Gorrell et al., 2018, pp. 601). Rather than using supervised classification techniques, they argue that their rule-based systems to detect abuse allows for a interpretable and easy to modify method, allowing researchers to address weaknesses of the approach without the need for additional large-scale quantities of data.[1] However, this approach is a laborious one as it requires the researchers to manually identify patterns of abuse and construct rules that can address such patterns along with any exceptions to the patterns that are not abusive.

**Prior work on lexical replacements** Reducing the dimensionality of data has been approached from a number of different avenues in prior literature. A large body of work is dedicated to the mathematical and algorithmic induction of what features to retain and which to omit (Witschel and Biemann, 2006, e.g.). The object of the mathematical and algorithmic approaches is to identify which features provide the most and least information about the distinctions between each class. Another body of work has been dedicated to dimensionality reduction through semantic replacements. One approach to such a reduction is the use of clustering algorithms, e.g. Brown Clustering (Derczynski et al., 2015). Another, more recent approach is to perform partial replacements in documents, only replacing some tokens. This method has been used in recent studies on abuse detection by using the HurtLex resource (Bassignana et al., 2018). HurtLex is a multi-lingual lexicon that seeks to map 'offensive' words into three macro categories, and seventeen fine-grained categories.

Several papers have approached the task of abuse detection across multiple languages by using HurtLex to perform a reduction in the feature space of offensive words in the lexicon. For instance, Pamungkas et al. (2020) use HurtLex to create a feature vector for each of the fine-grained categories to be used for optimisation and classification, as a way to separate out profanities and hateful terms. Chiril et al. (2019b) use HurtLex across English and French tweets to count the number of times each of the fine-grained categories are invoked. They optimise their models using the length of the tweets along with the number of times each fine-grained category is invoked in a tweet. Others have used HurtLex in combination with deep learning methods by creating onehot encodings (Pamungkas and Patti, 2019) and using the encoding to create a HurtLex embedding layer (Koufakou et al., 2020). These approaches all either directly reduce the feature space or seek to force optimised models to pay particular attention to the tokens included in the HurtLex resource. The approach that I take in this chapter contrasts these approaches by replacing all tokens with their respective LIWC categories and optimising models on the resulting document representations.

In this chapter, I take inspiration from the use of manually crafted features as a way to provide testable hypothesis while departing from the notion of feature generation. Specifically, I

[1] This detail on the rule-based nature of the classification systems was provided by Genevieve Gorrell in personal communications.

hypothesise that LIWC categories can provide information for predictive modelling that can allow for high performance in spite of token sparsity when using neural network methods.

#### 5.1.2.2 Neural Networks

Though the earliest models for the tasks were predominately linear models that used manually generated features (Davidson et al., 2017; Talat and Hovy, 2016; Warner and Hirschberg, 2012) more recent work has been dominated by the development of neural network-based models for automated abuse detection, posting ever-evolving state-of-the-art models and classification performances (Badjatiya et al., 2017; Isaksen and Gambäck, 2020; Park and Fung, 2017; Stoop et al., 2019; Zimmerman et al., 2018, e.g.). Here I consider a handful of neural network methods for detecting abuse, focusing on the distinct implications following the modelling choices and the logics that underpin them. As all neural network-based methods that I examine receive only the text as input, the primary differences between different proposed neural network models is in their use and organisation of different types of layers and the loss function selected for the respective models.

The most commonly used architecture for neural networks that in the surveyed literature is a CNN (Gambäck and Sikdar, 2017; Kolhatkar et al., 2020; Park and Fung, 2017; Wang et al., 2020; Wulczyn et al., 2017; Zimmerman et al., 2018). As CNNs have been the subject of particularly interest, a number of distinct modelling approaches have been proposed. First, relying on a simple neural network architecture, Kolhatkar et al. (2020) use GloVe embeddings as the first layer, followed by three convolutional layers (that have window sizes $3,4$, and 5, respectively) with global maximum pooling layers. Prior to passing data to an output layer, a dropout layer is applied to the output of the convolutional layers which is then passed to a dense layer. All layers prior to the output layer use a ReLU (see chapter 3 for more detail) activation function. The output layer applies the sigmoid function to provide a prediction from the model. This model most closely resembles the CNN architecture used in this chapter. As this model uses a pre-optimised word-embedding layer as its input layer, the input the model receives are documents that have been subject to tokenisation processes.

A different architecture is proposed by Park and Fung (2017). In their work they compare a single classifier, what they name a 'one-step classifier', that predicts the final classes directly with a stacked architecture of two models, with a 'two-step classifier' in their vernacular, that first predicts whether content is abusive and then predicts which type of abuse the documents predicted as abusive are. There are two primary distinctions between the two-step architecture proposed by Park and Fung (2017) and the architecture proposed by Kolhatkar et al. (2020). First, Kolhatkar et al. (2020) acts as a one-step classifier whereas the architecture proposed by Park and Fung (2017) acts in two steps. Second, Kolhatkar et al. (2020) only acts on documents tokenised into words and punctuation whereas Park and Fung (2017) propose a CNN that takes documents tokenised into words and punctuation in addition to documents tokenised into characters. Park and Fung (2017) show that through the use of a one-step CNN optimised on word and character input, they achieve a performance boost obtaining a *F1-score* of 0.827 on the datasets proposed by Talat and Hovy (2016) and Talat (2016), though the performance boost is lost once a two-step hybrid CNN is used.

As CNNs build feature mappings by passing over the data using filters, they come with certain assumptions built into them. As researchers define the number of filters and the stride size, they also define the range, in terms of token order, within which the model is given leave to identify token interactions. The implication of this is that there will likely be some, potentially overlapping ranges that the models identify patterns from. Depending on how researchers define these, the models will develop feature mappings corresponding to the ranges provided.

Another frequently used architecture is LSTMs (Badjatiya et al., 2017; Kolhatkar et al., 2020; Meyer and Gambäck, 2019). For instance, Kolhatkar et al. (2020) propose using a bi-directional LSTM that, like their CNN, has a pre-optimised embedding layer, a recurrent layer, a dropout layer, and a fully connected output layer with a sigmoid activation to predict the output classes. Meyer and Gambäck (2019) on the other hand take develop on the idea of a hybrid CNN, developing a LSTM architecture that takes documents tokenised into words and characters as input. The word representation is obtained through tokenisation passed through an embedding layer and the character representation is obtained by processing the documents with a CNN. Using these approaches, Kolhatkar et al. (2020) show comparable performances between the CNN and bi-directional LSTM on their dataset. Meyer and Gambäck (2019) on the other hand show that a baseline model only using character level information performs comparably with other more complex approaches, obtaining a macro *F1-score* of 0.7923 for the baseline and 0.7924 for the final system on the dataset proposed by Talat and Hovy (2016), and notably out-performs several other previously proposed methods.

The use of LSTMs, that rely on recurrence, breaks with the independence assumption of the manual feature-based models. By recurring over a document, each new token is considered in conjunction with the previous tokens that have not been forgotten. In this way, an assumption is built into the models that through processing enough token sequences, it will be possible to identify patterns in the sequences of tokens that connote abuse. Such a reliance on the text alone does not consider the positionality of abuse; Talat et al. (2018) argue only through understanding the context within which the speaker and audience exist in, is it possible to deem something as abusive. For instance, it is only through an understanding of the speaker that one can deem whether the *n-word* is weaponised as abuse or is reclaimed to connote complex social identity.

All methods described that rely on documents tokenised into sentences rely on pre-optimised embedding layers (most frequently GloVe Pennington et al. (2014)), come with their own costs and benefits. For instance, word embeddings that are optimised on web-text are likely to harbour social biases (Bolukbasi et al., 2016) that have been proven hard to address (Gonen and Goldberg, 2019). On the other hand, they also allow for better representations of related concepts and will be less susceptible to creating different representations for closely related concepts as a result of dataset biases. For instance, the concepts 'Television' and 'T.V.' might only be distantly related, if at all, in a small dataset due to few co-occurring terms within the dataset. In a larger dataset, spanning millions of documents, these two concepts are likely to appear as closely related as a robust language representation will likely have been achieved for such commonly occurring tokens. The methods that rely on character embeddings are also subject to similar distributional concerns, however this can be a benefit when used in conjunction with word embeddings. This benefit comes through as the set of possible characters is much smaller than the set of possible unique words, less data is needed to optimise robust embedding layers, though the optimised character embeddings will be particularly attuned to the dataset at hand. On the other hand, due to such particularity of the character embeddings, they are less likely to map well onto other domains even if they show good performance on the dataset that they are derived from.

For the work in this chapter, the use of pre-optimised embeddings is not appropriate for some models. Specifically the models that use LIWC-represented documents as LIWC embeddings are not publicly available or have been developed, to the best of my knowledge. Moreover, documents represented through LIWC categories are poorly suited for optimising general embeddings as only a small set of tokens are defined and they are not necessarily distributed in a suitable fashion for developing such generalised embeddings. Second, I don't use pre-optimised embeddings in the architectures for other models as a means to ensure that any comparisons with the LIWC-based models, provide a direct comparison of the influence of using LIWC as input tokenswhile avoiding potentially confounding factors.

### 5.1.3 Datasets

In order to understand and validate my approach, I optimise a model on multiple datasets. Moreover, I take each model that is optimised on a given dataset and apply it to all other datasets. To accommodate prediction on a model optimised on one dataset to others, I reduce all classification tasks to a binary task of abusive and not-abusive. This has downstream implications for the construction of the datasets and for the validity of the prediction task on the out-of-domain datasets. First, the dataset distributions are modified as tasks with more than two classes see their data collapsed. For some datasets, this means that the class imbalances are improved, as the majority class is non-abusive. The exception to this is the dataset proposed by Davidson et al. (2017) where the largest class is the 'offensive' class, which I combine with 'hateful', further minimising the size of the negative class. Second, as each dataset has been collected with different rationales and annotated with distinct purposes (please see section 3.1.2.1 for more detail), direct comparisons, and subsequently model predictions on each dataset, can be at odds with the goals of the datasets. For this reason, high scores on prediction metrics on external datasets should be viewed as a weak indication of the ability to identify general patterns while low scores can indicate a number of factors including, but not limited to, highly distinct data sources, annotation strategies, and lastly the questions each dataset is developed to ask.

With these concerns in mind, I decide to use datasets with distinct sources that are developed for different purposes. Rather than resist or seek to minimise the modelling concerns, I choose to lean into them to allow space for understanding how LIWC-based modelling may influence the optimisation and model performance on each dataset. Moreover, I seek to gain an understanding along which axes model generalisation may be afforded using LIWC-based modelling (see table 5.1 for the vocabulary sizes for each dataset and input type).

In this chapter I use the *StormFront* dataset (Garcia et al., 2019), the *Offence* dataset (Davidson et al., 2017), the *Hate Speech* dataset (Talat and Hovy, 2016), the *Expert Hate* dataset (Talat, 2016), and finally the *Toxicity* dataset (Wulczyn et al., 2017) (please see section 3.1.2.1 for a detailed overview of each dataset).

#### 5.1.3.1 StormFront

First, I use the *StormFront* dataset which was collected from the white supremacist web forum of the same name by Garcia et al. (2019). The data consists of $2,392$ documents, split into $1,531$ optimisation documents, 383 documents for validation, and 478 test documents. While the full dataset published consists of $10,000$ documents annotated as 'hate' and 'not-hate', with a large class imbalance towards non-hateful comments, I choose to use a balanced subset of the data provided by the authors. I choose this subset as it allows for testing how LIWC-based models perform when optimised on a a) small dataset and b) balanced data distribution. The dataset is initially split into an optimisation and evaluation set, I further create a validation set by extracting a stratified sample from the optimisation data, retaining the class balance from the balanced subset.

#### 5.1.3.2 Offence

The second dataset that I use to optimise and evaluate my models is the *Offence* dataset collected from Twitter by Davidson et al. (2017). This dataset was collected to disambiguate offensive tweets from hateful ones. This dataset is distinguished from all other datasets in that the positive classes, i.e. 'offensive' and 'hateful' accounting for $1,430$ documents and $19,190$ documents, respectively. This leaves only $4,163$ documents in the negative class. Once binarised, the dataset consists of $4,163$ documents in the negative class and $20,620$ documents in the positive class. The dataset is provided by the authors as a single file containing all documents, so I create stratified splits of the data into an optimisation set (80% or $19,826$ documents), a validation set (10% or $2,478$ documents), and a evaluation set (10% or $2,479$ documents), retaining the class distribution in each split. Using this dataset further allows for an investigation into how sensitive LIWC-based modelling is to dataset skews.

#### 5.1.3.3 Hate Speech

I also use the *Hate Speech* dataset, which is collected from Twitter by Talat and Hovy (2016). This dataset contains $16,914$ documents that follow a more traditional class distribution for abusive language data. In this dataset I collapse the 'racism' and 'sexism' classes into a single positive class, 'abuse'. This class consists of $5,355$ documents with the negative class occupying the remaining $11,559$ documents. The primary function that this dataset serves in this chapter is to allow for insight into whether the LIWC-based models would function under a distinct annotation criteria that is motivated by academic work in Gender Studies and Critical Race Theory on marginalisation, rather than social media guidelines for acceptable behaviour.

#### 5.1.3.4 Expert Hate

The *Expert Hate* dataset proposed by Talat (2016) contains $6,909$ documents and is also collected from Twitter, and it is also designed as a multi-class classification task. In this dataset the positive classes consist of 'sexism' (13% or 898 documents), 'racism' (1.41% or 97 documents) and 'both' (0.70% or 48 documents) while the negative class consists of 84.19% of the dataset. I reduce this down to a binary classification task and split the dataset into an optimisation set (80% or $5,527$ documents), a validation set (10% or 690 documents), and an evaluation set (10% or 692 documents) ensuring that binary the class distribution is retained. This dataset is annotated following the annotation guidelines proposed by Talat and Hovy (2016), however it is annotated using intersectional feminist activists as annotators. This dataset then allows for testing the influence of LIWC-based models on data annotated by experts.

#### 5.1.3.5 Toxicity

Finally, I use the *Toxicity* dataset published by Wulczyn et al. (2017). This dataset was collected from Wikipedia editor discussion pages and annotated as 'toxic' and 'not-toxic', and it is the largest dataset with $159,686$ documents. These documents are provided split into an optimisation set consisting of $95,692$ documents, a validation set with $32,128$ documents and a evaluation set containing $31,866$ documents. Similarly to the *Hate Speech* and *Expert Hate* datasets, this dataset is highly imbalanced with the positive class accounting for $16\%$ of the entire dataset. I use this dataset to gain an understanding of how large scale datasets can influences the performance, size, and optimisation time of LIWC-based models.

## 5.2 Modelling

In order to understand the impact of LIWC-based modelling, I design feature-based and neural network-based machine learning models. I optimise a Logistic Regression model and a SVM model with a linear kernel for each type of input data (Word unigrams, BPE unigrams, and LIWC unigrams) to allow for feature-based analysis of the identified patterns. To investigate how neural network models operate on the input data, I develop three types of neural networks for each input type: First, I optimise a MLP to provide an initial insight into whether neural network approaches might be appropriate. Second, I develop a LSTM model to investigate whether there are any benefits from its recurrent nature. Lastly, I develop a CNN model due to their widespread use in previous work.

I specify two different optimisation procedures, one for the linear baseline models and one for the neural networks. For the linear baseline models, I tokenise and pre-process the data and perform a grid-search over a parameter space in search of the parameter values that optimises for the highest macro *F1-score* performance. For neural network models, I similarly tokenise and pre-process the data and perform a Bayesian hyper-parameter search to identify the best performing parameter setting given by macro *F1-score*. I then reuse this best performing parameter settings and re-run the models with 5 different random seeds to ensure that the models' behaviour on the dataset is not the result of the initialisation of tensors caused by the setting of the random seed values.

| Dataset | Word Vocabulary | BPE Vocabulary | LIWC Vocabulary |
|---|---|---|---|
| *Offence* | 16,768 | 16,663 | 857 |
| *Toxicity* | 95,710 | 95,712 | 1,024 |
| *Hate Expert* | 9,110 | 9,181 | 744 |
| *Hate Speech* | 14,730 | 14,834 | 849 |
| *StormFront* | 5,566 | 5,510 | 622 |

**Table 5.1** Vocabulary sizes for each input type and the optimisation set for each dataset.

### 5.2.1 Pre-processing

Prior to providing data to any model, it is necessary to pre-process the data to make it suitable for the experiment conducted. In my experiments I examine the influences that vocabulary manipulation has on model design. To this effect, it is necessary to have some shared and distinct pre-processing steps for the datasets, depending on the experiment. For the shared pre-processing steps, I lower-case all documents, replace all usernames, that follow the Twitter standard of an '@' followed by a string, with a generic <*USER*> token, replace all website URLs with a generic <*URL*> token, and finally, replace all hashtags with a generic <*HASHTAG*> token. The resulting vocabulary sizes for each dataset and data type can be seen in table 5.1.

For the LIWC-based models and the word-based models, I pre-process documents using the python library Ekphrasis (Baziotis et al., 2017) which was developed specifically to handle the particularities of social media text. For instance, elongated words are mapped to their unelongated form, e.g. 'heyyyy' is mapped to 'hey' (see table 5.2 and table 5.3 for examples of tokenisation). No further processing is done for experiments using word tokens as input.

For the LIWC experiments, I take another step after the initial tokenisation so that I can compute the LIWC categories invoked by each word. Each token obtained is passed through a function which identifies all LIWC categories that the token invokes and combines them into a single token, where each LIWC category is separated by an underscore. All tokens that are not recognised by LIWC are replaced with a general token for <*UNK*> token (see table 5.3 for examples on the result of the pre-processing of documents).

For the BPE-based models on the other hand, I pre-process documents using the 200-dimensional Byte-Pair Embeddings from the BPE python library (Heinzerling and Strube, 2018). Byte-Pair Encodings are well suited to handle the particularities of social media text, as it breaks unrecognised words into known subwords, thus minimising unknown tokens in the validation and evaluation sets. Through this process, the hope is that even if part of a of

| Document | Word Token Representation | Byte-Pair Representation |
|---|---|---|
| Man I fucking hate animals! | Man I fucking hate animals ! | _man _i _fucking _hate _animals ! |
| Man I fking h8 animals! | Man I fking h8 animals ! | _man _i _f king _h 0 _animals ! |
| Bruv I fking hate animals! | Bruv I fking hate animals ! | _br uv _i _f king _hate _animals ! |

**Table 5.2** Word token and BPE representation.

| Document | LIWC Representation |
|---|---|
| Man I fucking hate animals | MALE_SOCIAL PPRON_FUNCTION_I_PRONOUN AFFECT_SEXUAL_BIO_INFORMAL_NEGEMO_ANGER_ADJ_SWEAR AFFECT_NEGEMO_ANGER_VERB_FOCUSPRESENT UNK UNK |
| Man I fking h8 animals | MALE_SOCIAL PPRON_FUNCTION_I_PRONOUN UNK NUM UNK UNK |
| Bruv I fking hate animals | UNK PPRON_FUNCTION_I_PRONOUN UNK AFFECT_NEGEMO_ANGER_VERB_FOCUSPRESENT UNK UNK |

**Table 5.3** Examples of LIWC representations.

a word is out-of-vocabulary for the model some of its subwords will be within a model's vocabulary, allowing the remaining subwords to be used for inference.

In reviewing table 5.1, it is clear that computing LIWC representations result in smaller vocabularies than the surface form counter-parts. It's further apparent that the BPE representations does not substantially alter the vocabulary sizes, with the exception of the *Hate Expert* and *Hate Speech* datasets, where there's a small increase in the vocabulary sizes. It is unsurprising that the vocabulary size would grow when using BPE, as subwords for all unrecognised tokens are computed using the Byte-Pair Encoding. That is, when there are words in the dataset that aren't recognised by the embedding, the words are split into their sub-parts thus increasing the overall size of the dataset.

For the documents represented through the LIWC categories that they invoke, I observe in table 5.1 a sharp decline in the sizes of the vocabularies. This is expected as the LIWC dictionary only encompasses a small number of words. Many words used in informal conversations on online platforms are likely to fall outside of those considered when developing the dictionary. Moreover, it is also not surprising that a drop would occur as many of the datasets are created and published after the creation of the LIWC dictionary and examine domains that are unlikely to be well represented within the LIWC dictionary. Consequently the vocabularies produced using the LIWC dictionary is subject to language drift in addition to domain shifts.

Considering tables 5.4 to 5.6 that display the numbers of tokens that are unique to- and shared by each class in each dataset, there are some clear implications for my research question. First, only minor distributional shifts occur between word-based vocabularies (see table 5.5)

| | Not Abuse Only | Abuse Only | Intersection | Vocab size |
|---|---|---|---|---|
| *Offence* | 24 (2.8%) | 150 (17.6%) | 677 (79.6%) | 851 |
| *Toxicity* | 131 (12.8%) | 5 (0.5%) | 886 (86.9%) | 1,022 |
| *Hate Expert* | 241 (32.6%) | 25 (3.4%) | 473 (64%) | 739 |
| *Hate Speech* | 116 (13.9%) | 47 (5.62%) | 674 (80.5%) | 837 |
| *StormFront* | 74 (11.9%) | 117 (18.8%) | 431 (69.3%) | 622 |

**Table 5.4** Number of unique LIWC tokens in each class for each dataset and the size of their intersection.

| | Not Abuse Only | Abuse Only | Intersection | Vocab size |
|---|---|---|---|---|
| *Offence* | 3,303 (19.7%) | 8,656 (51.6%) | 4,809 (28.7%) | 16,768 |
| *Toxicity* | 71,491 (74.7%) | 1,560 (1.6%) | 22,659 (23.7%) | 95,710 |
| *Hate Expert* | 6,155 (67.6%) | 953 (10.5%) | 2,002 (22.98%) | 9,110 |
| *Hate Speech* | 7,042 (47.8%) | 2,599 (17.6%) | 5,089 (34.6%) | 14,730 |
| *StormFront* | 1,834 (32.9%) | 2,273 (40.8%) | 1,459 (26.2%) | 5,566 |

**Table 5.5** Number of unique word tokens in each class for each dataset and the size of their intersection.

and BPE-based vocabularies (see table 5.6). As processing and representing documents as their byte-pair represented counter-parts results in the computation of sub-words, such small distributional discrepancies are to be expected. Second, observing the differences between LIWC vocabulary distributions (see table 5.4) and the word vocabulary distributions, it is clear that there are large distributional changes, which have large ramifications for the datasets and subsequently the models optimised on them. This means that as the vast majority of tokens are shared between the classes, there are fewer potential signals for models to over-fit to. Disregarding token interactions, a word-based model optimised on the *Toxicity* dataset may overfit to 72,929 unique tokens (76.3% of all unique tokens in the dataset). Similarly disregarding token interactions, a LIWC-based model is only provided with 136 unique tokens (13.3% of all unique tokens in the dataset) to which it can overfit to. Similarly to n-gram character-based modelling, a smaller set of unique tokens is likely to result in a matrix that is, in places, more dense, allowing for a model to identify patterns based on the interaction of tokens rather than individual tokens. This particular case is likely for LIWC-based models as the vast majority (between 64% and 86%) of tokens are shared between both classes.

### 5.2.2 Linear Baseline Models

For the linear baseline models, I optimise several different linear models (i.e. Logistic Regression models and SVM models) that function as baselines using the Scikit-Learn python library Pedregosa et al. (2011). For each algorithm, I develop three different models:

| | Not Abuse Only | Abuse Only | Intersection | Vocab size |
|---|---|---|---|---|
| *Offence* | 3,199 (19.2%) | 7,978 (47.9%) | 5,486 (32.9%) | 16,663 |
| *Toxicity* | 71,493 (74.7%) | 1,560 (1.6%) | 22,659 (23.4%) | 95,712 |
| *Hate Expert* | 6,231 (67.9%) | 971 (10.6%) | 1,979 (21.6%) | 9,181 |
| *Hate Speech* | 7,074 (47.7%) | 2,653 (17.9%) | 5,107 (34.4%) | 14,834 |
| *StormFront* | 1,804 (32.7%) | 2,240 (40.7%) | 1,466 (26.6%) | 5,510 |

**Table 5.6** Number of unique BPE tokens in each class for each dataset and the size of their intersection.

a) surface token-based model that uses sentences tokenised into words, b) models on the Byte-Pair encoded representation, and c) models that use the LIWC-based representation as their input data. For all baseline models, I only use token unigrams as features, as these provide competitive baselines for many of the datasets. To ensure that the baseline models use the most appropriate parameters, I perform a cross-validated grid-search (as implemented by Pedregosa et al. (2011)) over the possible settings of the model parameters for each model. For both SVM and Logistic Regression models, I explore values of the strength of the regularisation ($[0.1, 0.2, 0.3, \ldots, 0.9]$) to examine the strength of regularisation and the regulariser ($\{L1, L2\}$). For Logistic Regression, I also set the parameter search to consider *Elasticnet* as a third regulariser option.

In the optimisation procedure for the linear models, I first fit a count vectoriser on the optimisation data and then optimise a model on the vectorised optimisation data. For prediction on other datasets, all datasets are passed through vectoriser that is fitted to the optimisation data. This ensures that all datasets are processed and indexed in accordance to the vocabulary of the optimisation dataset and the model. A notable difference between the optimisation of linear models and their neural network counterparts is that linear models are only provided with the dataset once for each cross-validation set and the order of the documents in the dataset is not randomised whereas the neural network models iterate multiple times over the optimisation dataset where the order of the documents in the dataset is randomly shuffled between each iteration.

### 5.2.3 Neural Models

I implement three different neural network model types using PyTorch (Paszke et al., 2019) and perform a hyper-parameter search on each model type for every dataset and input type. Specifically, I implement a MLP, a LSTM, and a CNN. I choose to implement an MLP as it is the simplest form of neural networks and it can provide insights into the applicability of neural network-based architectures. As the LIWC tokens are distributed such that the vast majority of tokens are shared by both classes, I also develop a LSTM network to take long-range dependencies into account. I choose a LSTM over a basic RNN model as unknown tokens are likely to occur frequently in the LIWC-based data due to the small size of the dictionary and resulting vocabularies, and it may be desirable for any recurrent model optimised on the data to be afforded the ability to forget sequences of unknown tokens. Finally, I develop a CNN model as this model type has been frequently applied in the previous work.

In order to focus on the utility of the different document representations, I optimise models that have simple architectures. To this end, I don't use pre-optimised embeddings as embedding layers within the model as I am not aware of any general purpose pre-optimised LIWC-embeddings and, as is apparent from table 5.3, the LIWC tokens generated for each token would most likely be out-of-vocabulary for most pre-optimised word embeddings. Instead, I opt to optimise the embedding layer along with all other layers. To address the issue of the model over-fitting the data, either by identifying spurious correlations in the data or by over-optimising the model, I subject each model to dropout and early stopping (see section 3.2 for more detail on dropout and early stopping). To address the issues of exploding and vanishing gradients, I employ gradient clipping (Bengio et al., 1994), to normalise the value of the gradients in the optimisation procedures.

I use a single optimisation procedure for all models to limit the confounding factors in the optimisation process. The models are given data, which they iterate over in a pre-defined number of epochs, shuffling the dataset between each epoch. Within each epoch, batches of the data are passed through the model for prediction during optimisation. Following this prediction, the loss is back-propagated through the model, updating the internal representation in the process. This process is repeated for the assigned maximum number of epochs, or until the model triggers the early stopping (Prechelt, 1998) criteria. The early stopping criteria is that the computed loss on the validation set has been strictly increasing for at least 15 epochs. Once a model has finished optimising, performance in terms of macro `F1-score`, `precision`, `recall` and `accuracy` are computed on the validation set and evaluation sets. To be able to speak to the optimisation time, I start a timer when the model optimisation procedure is initiated and stop the timer when the model has fully completed its optimisation, but prior to any inference made using the model. I repeat this process for at least 200 unique trials for each model and dataset combination and use the `F1-score` on the validation set to identify the best configuration of hyper-parameters. Once the best hyper-parameters are selected, I rerun the models with 5 different random seeds and obtain these models' performance on all evaluation sets, that is the in-domain test for the dataset the models are optimised on and the out-of-domain evaluation sets from the remaining datasets.

In efforts to identify the best hyper-parameters without performing a grid-search of all possible combinations, I turn towards Bayesian Hyper-Parameter Tuning (Neal, 1996). Briefly, Bayesian Optimisation allows for estimating the best hyper-parameters for a model through a series of trials with different hyper-parameter settings. I use the implementation of Bayesian Hyper-Parameter Optimisation offered through Biewald (2020) and set the objective of the hyper-parameter optimisation to maximise the macro `F1-score` on the development data (please refer to section 3.2.4 for more detail).

The parameters that I perform the optimisation for varies across the different model types as they require different hyper-parameters to be defined. A set of hyper-parameters are constant across models: the size of mini-batches provided to the model for optimisation, the learning rate, the number of epochs, and the embedding size. For each dataset and model type, I perform at last 200 trials with different parameter settings, leading to choosing a final set of hyper-parameters that I run with five different random seeds. The values for the learning rate are sampled from a uniform distribution while the batch size and number of epochs to optimise for are sampled from a pre-defined categorical set. More generally, the values for all hyper-parameters, asides from dropout and the learning rate are sampled from a pre-defined categorical set.

- Maximum epoch count: $\{50, 100, 200\}$,
- Batch size: $\{16, 32, 64\}$,
- learning rate: $[0.00001, 1.0]$

#### 5.2.3.1 Multi-Layered Perceptron

The first neural architecture that I implement is a Multi-Layered Perceptron. I choose this model as it is the simplest form of a neural network and thus well suited for an initial investigation into the feasibility of neural network approaches for the LIWC-based representations. The MLP architecture that I use also lays the basis for the architectures for all other neural network-based models in this chapter. The network consists of an embedding input layer, a hidden layer, an output layer, and a softmax layer which produces the probabilities for each class. I subject the model's representation to a dropout layer and a non-linear activation function between the input layer and hidden layer, and the hidden and output layer.

I perform a hyper-parameter search over the following hyper-parameter values, that are specific to the MLP:

- dropout probability: $[0.0, 0.5]$,

- hidden layer dimension: $\{64, 100, 200, 300\}$, and
- the activation function: $\{ReLU, Tanh\}$

#### 5.2.3.2 Long-Short Term Memory

I implement a simple architecture for the LSTM model to retain focus on the input types and to avoid differences in model architectures, such as the use of an attention layer, to ensure that the I examine the impact of the data representation, rather than potential differences the construction of the models. The LSTM consists of an input embedding layer, a unidirectional LSTM layer, a linear output layer, and a softmax to produce the class probability distribution. I apply a dropout to the output of the input layer and the output of the LSTM layer in efforts to prevent the model from over-fitting on any particular pattern. I use the PyTorch implementation of an LSTM which uses a Tanh activation function (Paszke et al., 2019), for which reason I do not apply other non-linearities to the model.

Thus, for the LSTM our hyper-parameter tuning considers the following parameters and values:

- dropout probability: $[0.0, 0.5]$,
- embedding layer dimension: $\{64, 100, 200, 300\}$, and
- hidden layer dimension: $\{64, 100, 200, 300\}$

#### 5.2.3.3 Convolutional Neural Network

Similarly to the MLP and LSTM, the input layer is an embedding layer, followed by three two-dimensional convolutional layers, each of which are subject to a non-linear activation function, a one dimensional max-pooling layer, and a output layer. The representation obtained through the output layer is subjected to a softmax function that computes the probability distribution for the classes.

Unlike the MLP and LSTM models, the CNN models are not subject to a dropout. However as the CNN requires window sizes and number of filters to be set, these are added as hyper-parameters to tune. The hyper-parameters that are used to tune this model are thus:

- window size: $\{(1, 2, 3), (2, 3, 4), (3, 4, 5)\}$,
- Number of filters: $\{64, 128, 256\}$,
- hidden layer dimension: $\{64, 100, 200, 300\}$, and

- the activation function: $\{ReLU, Tanh\}$

## 5.3 Results

To answer the research question set forth, I examine three different aspects of my models: First, I examine the model performances on a held out evaluation set from dataset they are optimised on in an effort to answer how appropriate LIWC representations are for automated abuse detection. Second, I compare the model performances on the evaluation sets of datasets that they are not optimised on. Finally, I observe the time it takes for models to be optimised using the different representations. To examine these behaviours across datasets and architectures, I develop three different neural network types and optimise these on three different data representations for each of the five datasets introduced in section 5.1.3, resulting in 45 different model architectures optimised. These 45 model architectures are then subject to at least 200 hyper-parameter selection trials, resulting in over $9,000$ models developed in the process of identifying the best hyper-parameters. Once the hyper-parameters are determined, I perform an additional 5 runs for each model architecture, examining the influence of the random seed.

### 5.3.1 Baseline Models

#### 5.3.1.1 Model Parameters and Validation Set Performances

All linear baseline models prefer an L2 regularisation. Considering table 5.7, it is clear that in most cases using a word token input results in the best scores on the development set, the exception to the rule being the *Hate Expert* dataset. However, an interesting pattern emerges, for many of the datasets, using the LIWC-tokenised input provides highly competitive results, suggesting the efficacy of using LIWC-based tokenisation even on linear models, a promising sign for the subsequent experiments.

#### 5.3.1.2 Evaluation Set Performances

In tables 5.8 to 5.12, though only baselines, there are a number of interesting patterns that emerge. First, the baselines validate the hypothesis that LIWC-based representation can serve as a viable input to machine learning models, as the LIWC-based models often achieve competing scores, and in some instances out-perform models with other data representations, e.g. in `recall` on the in-domain prediction on the *Offence* test data (see table 5.8) and the

| | | Model | C | F1-score |
|---|---|---|---|---|
| Offence | Word | SVM | 0.1 | **0.9222** |
| | | LR | 0.9 | 0.9093 |
| | BPE | SVM | 0.1 | 0.9216 |
| | | LR | 0.8 | 0.9119 |
| | LIWC | SVM | 0.1 | 0.9207 |
| | | LR | 0.2 | 0.9140 |
| Toxicity | Word | SVM | 0.2 | **0.8678** |
| | | LR | 1.0 | 0.8660 |
| | BPE | SVM | 0.1 | 0.8664 |
| | | LR | 1.0 | 0.8673 |
| | LIWC | SVM | 0.9 | 0.8514 |
| | | LR | 1.0 | 0.8374 |
| Hate Expert | Word | SVM | 0.1 | 0.7587 |
| | | LR | 1.0 | 0.7653 |
| | BPE | SVM | 0.1 | **0.8090** |
| | | LR | 0.8 | 0.7974 |
| | LIWC | SVM | 1.0 | 0.6378 |
| | | LR | 0.7 | 0.6354 |
| Hate Speech | Word | SVM | 0.1 | **0.7995** |
| | | LR | 0.9 | 0.7928 |
| | BPE | SVM | 0.1 | 0.7853 |
| | | LR | 0.5 | 0.7676 |
| | LIWC | SVM | 0.4 | 0.7214 |
| | | LR | 1.0 | 0.7265 |
| StormFront | Word | SVM | 0.1 | 0.7485 |
| | | LR | 0.9 | **0.7508** |
| | BPE | SVM | 0.3 | 0.7041 |
| | | LR | 1.0 | 0.7406 |
| | LIWC | SVM | 0.1 | 0.7068 |
| | | LR | 0.1 | 0.7249 |

**Table 5.7** Optimal parameter values for linear baseline models optimised on each data set and their in-domain performance on the validation sets.

`F1-score` achieved on the out-of-domain datasets, for instance for the models optimised on the *Toxicity* dataset (see table 5.9).

Second, though the baselines often produce sub par classification performances on out-of-domain data, ranging from random performance to worse than a majority baseline, they sometimes yield surprisingly good results along individual metrics. E.g. in table 5.10 a word token model achieves a high performance on the *Toxicity* dataset while a LIWC-based model posts surprisingly high scores on the *Offence* dataset.

Third, many of the LIWC-based models post strong performances on in-domain data, suggesting that simple linear models may be highly appropriate for the LIWC-based input. Moreover, these strong in-domain performances also provide a weak suggestion that neural network architectures may be well suited for the task, as most of the of the datasets that are under consideration are small datasets. Fourth, an interesting trend of the models optimised on the *Offence* and the *Toxicity* datasets obtain surprisingly good scores on each other. Two potential reasons for this trend are, 1) the two datasets are the largest in datasets so more general patterns may be learned, 2) as the notion of 'offensive' as constructed by Davidson et al. (2017) bears strong similarities with the notion of 'toxicity' constructed by Wulczyn et al. (2017), they may yield subsets of the datasets that strongly share similarities with each other.

Finally, although these baseline models do obtain surprisingly good scores for many of the models, there are several instances of noteworthy performance drops between the word token and BPE-token models and their LIWC counterpart, notably the LIWC-based models often perform well on some metrics, but fall short on others. For instance in tables 5.11 and 5.12, the LIWC-based models perform well on `recall` and `precision`, though they are frequently out-performed by models optimised with different data representations. Thus, within the space of these strong results there is still room for improvement of the metrics. As most of the datasets are imbalanced, I focus my attention on the macro `F1-score` performance due to its particular ability to handle imbalanced data well and its use in the previous literature (Fortuna et al., 2021).

### 5.3.2 Neural models

### 5.3.3 Model Hyper-Parameters and Validation Set Performances

As a result of the 200 trials, several parameter settings compete to be the best performing model, with little difference in their scores on the validation set. The hyper-parameters for the best and most stably performing model are presented in tables 5.13 to 5.17.

| | | Word | | BPE | | LIWC | |
|---|---|---|---|---|---|---|---|
| | | LR | SVM | LR | SVM | LR | SVM |
| *Offence* | Accuracy | 0.9504 | **0.9544** | 0.9484 | 0.9504 | 0.9399 | 0.9407 |
| | Precision | 0.9098 | **0.9140** | 0.9064 | 0.9073 | 0.8772 | 0.8755 |
| | Recall | 0.9137 | 0.9257 | 0.9095 | 0.9174 | 0.9275 | **0.9385** |
| | F1-score | 0.9118 | **0.9197** | 0.9079 | 0.9123 | 0.8995 | 0.9025 |
| Toxicity | Accuracy | 0.8319 | 0.8311 | 0.8882 | 0.8904 | **0.9342** | 0.9308 |
| | Precision | 0.6505 | 0.6467 | 0.7055 | 0.7077 | **0.8108** | 0.7988 |
| | Recall | 0.7941 | 0.7828 | 0.8006 | 0.7949 | 0.8062 | **0.8067** |
| | F1-score | 0.6800 | 0.6752 | 0.7393 | 0.7397 | **0.8085** | 0.8027 |
| Hate Expert | Accuracy | 0.6416 | 0.6792 | 0.6922 | 0.7413 | 0.7630 | **0.7688** |
| | Precision | 0.4892 | 0.4927 | 0.5223 | 0.5365 | **0.5738** | 0.5685 |
| | Recall | 0.4832 | 0.4899 | 0.5316 | 0.5416 | **0.5857** | 0.5737 |
| | F1-score | 0.4748 | 0.4864 | 0.5193 | 0.5381 | **0.5783** | 0.5708 |
| Hate Speech | Accuracy | 0.6413 | 0.6519 | 0.6294 | 0.6554 | **0.6738** | 0.6696 |
| | Precision | 0.5835 | 0.5834 | 0.5565 | 0.5778 | **0.5941** | 0.5838 |
| | Recall | **0.5810** | 0.5722 | 0.5503 | 0.5601 | 0.5564 | 0.5475 |
| | F1-score | **0.5820** | 0.5741 | 0.5510 | 0.5597 | 0.5490 | 0.5364 |
| StormFront | Accuracy | 0.5879 | 0.5921 | 0.5795 | **0.6172** | 0.5000 | 0.5314 |
| | Precision | 0.5887 | 0.5946 | 0.5798 | **0.6181** | 0.5000 | 0.5705 |
| | Recall | 0.5879 | 0.5921 | 0.5795 | **0.6172** | 0.5000 | 0.5314 |
| | F1-score | 0.5869 | 0.5893 | 0.5791 | **0.6164** | 0.4226 | 0.4559 |

**Table 5.8** Scores on the evaluation sets for linear models optimised on the *Offence* dataset.

| | | Word | | BPE | | LIWC | |
|---|---|---|---|---|---|---|---|
| | | LR | SVM | LR | SVM | LR | SVM |
| Offence | Accuracy | 0.6414 | 0.7019 | 0.6563 | 0.6962 | 0.8072 | **0.8665** |
| | Precision | 0.6442 | 0.6609 | 0.6501 | 0.6636 | 0.7291 | **0.7763** |
| | Recall | 0.7576 | 0.7825 | 0.7686 | 0.7897 | 0.8736 | **0.9083** |
| | F1-score | 0.5983 | 0.6459 | 0.6109 | 0.6437 | 0.7496 | **0.8116** |
| *Toxicity* | Accuracy | 0.9577 | **0.9577** | 0.9584 | 0.9583 | 0.9530 | 0.9549 |
| | Precision | 0.9058 | 0.8961 | 0.9052 | 0.9015 | 0.9135 | **0.9210** |
| | Recall | 0.8352 | **0.8478** | 0.8406 | 0.8446 | 0.7952 | 0.8008 |
| | F1-score | 0.8662 | 0.8700 | 0.8693 | **0.8703** | 0.8420 | 0.8484 |
| Hate Expert | Accuracy | 0.8020 | 0.8035 | **0.8353** | 0.8309 | 0.8309 | 0.8309 |
| | Precision | 0.4834 | 0.5111 | **0.6325** | 0.6131 | 0.6167 | 0.6262 |
| | Recall | 0.4929 | 0.5053 | 0.5513 | 0.5449 | 0.5524 | **0.5640** |
| | F1-score | 0.4786 | 0.4974 | 0.5582 | 0.5493 | 0.5600 | **0.5750** |
| Hate Speech | Accuracy | 0.6761 | 0.6767 | 0.6684 | 0.6690 | 0.6785 | **0.6838** |
| | Precision | 0.5876 | 0.5916 | 0.5606 | 0.5650 | 0.5925 | **0.6099** |
| | Recall | 0.5325 | 0.5378 | 0.5217 | 0.5246 | 0.5325 | **0.5423** |
| | F1-score | 0.5023 | 0.5130 | 0.4866 | 0.4925 | 0.5005 | **0.5167** |
| StormFront | Accuracy | 0.5544 | 0.5649 | **0.5649** | 0.5628 | 0.5126 | 0.5146 |
| | Precision | 0.6697 | 0.6645 | **0.6829** | 0.6675 | 0.5866 | 0.5799 |
| | Recall | 0.5544 | 0.5649 | **0.5649** | 0.5628 | 0.5126 | 0.5146 |
| | F1-score | 0.4632 | **0.4872** | 0.4811 | 0.4817 | 0.3800 | 0.3901 |

**Table 5.9** Scores on the evaluation sets for linear models optimised on the *Toxicity* dataset.

| | | Word | | BPE | | LIWC | |
|---|---|---|---|---|---|---|---|
| | | LR | SVM | LR | SVM | LR | SVM |
| Offence | Accuracy | 0.3651 | 0.3836 | 0.2925 | 0.3066 | 0.6006 | **0.6575** |
| | Precision | 0.5537 | 0.5551 | 0.5495 | 0.5511 | 0.6213 | **0.6317** |
| | Recall | 0.5695 | 0.5759 | 0.5451 | 0.5507 | 0.7159 | **0.7339** |
| | F1-score | 0.3620 | 0.3783 | 0.2922 | 0.3066 | 0.5608 | **0.6025** |
| Toxicity | Accuracy | **0.9054** | 0.9029 | 0.9022 | 0.9004 | 0.8482 | 0.8419 |
| | Precision | **0.7241** | 0.6800 | 0.6698 | 0.6525 | 0.6011 | 0.5973 |
| | Recall | 0.5339 | 0.5299 | 0.5291 | 0.5323 | 0.6265 | **0.6293** |
| | F1-score | 0.5404 | 0.5338 | 0.5325 | 0.5381 | **0.6111** | 0.6090 |
| *Hate Expert* | Accuracy | 0.8960 | **0.8988** | 0.8960 | 0.8974 | 0.8671 | 0.8584 |
| | Precision | 0.8162 | 0.8193 | **0.8327** | 0.8270 | 0.7537 | 0.7249 |
| | Recall | **0.7492** | 0.7625 | 0.7285 | 0.7446 | 0.6549 | 0.6536 |
| | F1-score | 0.7764 | **0.7865** | 0.7657 | 0.7765 | 0.6851 | 0.6779 |
| Hate Speech | Accuracy | 0.6885 | 0.6885 | 0.6791 | 0.6850 | **0.7063** | 0.7009 |
| | Precision | 0.6463 | 0.6463 | 0.5866 | 0.6199 | **0.6658** | 0.6474 |
| | Recall | 0.5285 | 0.5285 | 0.5158 | 0.5260 | 0.5774 | **0.5828** |
| | F1-score | 0.4777 | 0.4777 | 0.4577 | 0.4770 | 0.5681 | **0.5792** |
| StormFront | Accuracy | 0.5063 | 0.5063 | 0.5000 | 0.4979 | 0.5418 | **0.5460** |
| | Precision | **0.6516** | 0.6087 | 0.5000 | 0.2495 | 0.6364 | 0.6048 |
| | Recall | 0.5063 | 0.5063 | 0.5000 | 0.4979 | 0.5418 | **0.5460** |
| | F1-score | 0.3507 | 0.3541 | 0.3370 | 0.3324 | 0.4458 | **0.4720** |

**Table 5.10** Scores on the evaluation sets for linear models optimised on the *Hate Expert* dataset.

| | | Word | | BPE | | LIWC | |
|---|---|---|---|---|---|---|---|
| | | LR | SVM | LR | SVM | LR | SVM |
| Offence | Accuracy | 0.4490 | 0.4748 | 0.2727 | 0.4986 | **0.5450** | 0.5365 |
| | Precision | 0.5887 | 0.5926 | 0.5483 | **0.5955** | 0.6083 | 0.6073 |
| | Recall | 0.6353 | 0.6479 | 0.5381 | 0.6584 | 0.6873 | **0.6841** |
| | F1-score | 0.4378 | 0.4593 | 0.2714 | 0.4784 | **0.5168** | 0.5102 |
| Toxicity | Accuracy | **0.8786** | 0.8654 | 0.9006 | 0.8711 | 0.7278 | 0.7277 |
| | Precision | 0.5730 | 0.5571 | **0.6462** | 0.5788 | 0.5325 | 0.5333 |
| | Recall | 0.5390 | 0.5404 | 0.5268 | 0.5551 | 0.5702 | **0.5719** |
| | F1-score | **0.5463** | 0.5456 | 0.5290 | 0.5630 | 0.5222 | 0.5230 |
| Hate Expert | Accuracy | 0.8468 | 0.8555 | **0.8931** | 0.8540 | 0.8309 | 0.8324 |
| | Precision | 0.6881 | 0.7187 | **0.8246** | 0.7167 | 0.6447 | 0.6443 |
| | Recall | 0.6043 | 0.6172 | **0.7229** | 0.6273 | 0.5949 | 0.5881 |
| | F1-score | 0.6254 | 0.6428 | **0.7591** | 0.6525 | 0.6097 | 0.6030 |
| *Hate Speech* | Accuracy | **0.8381** | 0.8422 | 0.6838 | 0.8191 | 0.7719 | 0.7701 |
| | Precision | 0.8280 | **0.8332** | 0.6187 | 0.7968 | 0.7464 | 0.7446 |
| | Recall | 0.7882 | **0.7932** | 0.5193 | 0.7769 | 0.7006 | 0.6973 |
| | F1-score | 0.8030 | **0.8081** | 0.4602 | 0.7853 | 0.7138 | 0.7106 |
| StormFront | Accuracy | 0.5544 | **0.5816** | 0.5000 | 0.5565 | 0.5523 | 0.5586 |
| | Precision | 0.6858 | **0.7069** | 0.5000 | 0.6593 | 0.5929 | 0.6051 |
| | Recall | 0.5544 | **0.5816** | 0.5000 | 0.5565 | 0.5523 | 0.5586 |
| | F1-score | 0.4587 | **0.5069** | 0.3370 | 0.4712 | 0.4974 | 0.5037 |

**Table 5.11** Scores on the evaluation sets for linear models optimised on the *Hate Speech* dataset.

| | | Word | | BPE | | LIWC | |
|---|---|---|---|---|---|---|---|
| | | LR | SVM | LR | SVM | LR | SVM |
| Offence | Accuracy | 0.2929 | 0.2945 | 0.2912 | 0.3142 | 0.4248 | **0.6821** |
| | Precision | 0.5308 | 0.5372 | 0.5392 | 0.5398 | 0.5405 | 0.**6185** |
| | Recall | 0.5300 | 0.5357 | 0.5367 | 0.5429 | 0.5643 | **0.7027** |
| | F1-score | 0.2928 | 0.2944 | 0.2911 | 0.3141 | 0.4088 | **0.6078** |
| Toxicity | Accuracy | 0.6271 | **0.6393** | 0.6370 | 0.6233 | 0.4075 | 0.4606 |
| | Precision | 0.5085 | 0.5129 | 0.4987 | 0.5001 | 0.4959 | **0.5208** |
| | Recall | 0.5224 | 0.5336 | 0.4967 | 0.5002 | 0.4886 | **0.5587** |
| | F1-score | 0.4637 | **0.4726** | 0.4576 | 0.4541 | 0.3511 | 0.3944 |
| Hate Expert | Accuracy | 0.7731 | 0.7702 | **0.7905** | 0.7789 | 0.6792 | 0.6734 |
| | Precision | **0.5181** | 0.5111 | 0.5157 | 0.4923 | 0.4983 | 0.5091 |
| | Recall | **0.5144** | 0.5089 | 0.5095 | 0.4950 | 0.4976 | 0.5135 |
| | F1-score | **0.5146** | 0.5085 | 0.5063 | 0.4896 | 0.4921 | 0.5020 |
| Hate Speech | Accuracy | 0.6696 | 0.6732 | **0.6820** | 0.6803 | 0.6389 | 0.6294 |
| | Precision | 0.5821 | 0.5888 | **0.6081** | 0.6032 | 0.5928 | 0.5920 |
| | Recall | 0.5438 | 0.5459 | 0.5586 | 0.5524 | 0.5969 | **0.5998** |
| | F1-score | 0.5297 | 0.5312 | 0.5486 | 0.5388 | **0.5943** | 0.5932 |
| *StormFront* | Accuracy | 0.7259 | **0.7427** | 0.7280 | 0.7197 | 0.6987 | 0.6820 |
| | Precision | 0.7260 | **0.7428** | 0.7280 | 0.7201 | 0.7015 | 0.6830 |
| | Recall | 0.7259 | **0.7427** | 0.7280 | 0.7197 | 0.6987 | 0.6820 |
| | F1-score | 0.7259 | **0.7426** | 0.7280 | 0.7195 | 0.6977 | 0.6816 |

**Table 5.12** Scores on the evaluation sets for linear models optimised on the *StormFront* dataset.

Taking the best performing hyper-parameters, I re-run each model and dataset combination with five different random seeds (22, 32, 42, 310, and 922) and display the macro *F1-scores* and the loss on the in-domain validation sets in figs. 5.1 to 5.10. Note, that in figs. 5.1, 5.3, 5.5, 5.7 and 5.9 the standard deviations is depicted through shading whereas in figs. 5.2, 5.4, 5.6, 5.8 and 5.10 the shading represents the standard error.[2] Moreover, as loss values most frequently are small, I display the losses on a logarithmic scale to allow for more readable figures.

[2] I display the standard error here, rather than the standard deviation, as displaying the standard deviation on a log scale results in unreadable graphs due to outliers. Ignoring outliers in producing the graph has the natural consequence that parts of the graph exist outside of the displayed area.

| | Model | Embedding Dim | Hidden Dim | Window Size | Filters | Batch Size | Learning Rate | Dropout | Non-linearity | # Epochs |
|---|---|---|---|---|---|---|---|---|---|---|
| *WORD* | MLP | 200 | 100 | N/A | N/A | 32 | 0.8581 | 0.2434 | ReLU | 50 |
| | CNN | 100 | N/A | 1, 2, 3 | 256 | 32 | 0.0.1653 | N/A | ReLU | 50 |
| | LSTM | 64 | 200 | N/A | N/A | 64 | 0.0174 | 0.08569 | Tanh | 50 |
| *BPE* | MLP | 100 | 300 | N/A | N/A | 64 | 0.004817 | 0.2367 | ReLU | 200 |
| | CNN | 64 | N/A | 1, 2, 3 | 128 | 16 | 0.007243 | N/A | Tanh | 50 |
| | LSTM | 200 | 200 | N/A | N/A | 16 | 0.002144 | 0.05892 | Tanh | 50 |
| *LIWC* | MLP | 300 | 300 | N/A | N/A | 64 | 0.009468 | 0.2091 | ReLU | 200 |
| | CNN | 200 | N/A | 3, 4, 5 | 256 | 64 | 0.1044 | N/A | Tanh | 100 |
| | LSTM | 200 | 200 | N/A | N/A | 16 | 0.08958 | 0.4237 | Tanh | 50 |

**Table 5.13** Best hyper-parameters for models optimised on the *Offence* dataset.

| | Model | Embedding Dim | Hidden Dim | Window Size | Filters | Batch Size | Learning Rate | Dropout | Non-linearity | # Epochs |
|---|---|---|---|---|---|---|---|---|---|---|
| *WORD* | MLP | 200 | 100 | N/A | N/A | 64 | 0.001493 | 0.1396 | ReLU | 50 |
| | CNN | 64 | N/A | 3, 4, 5 | 128 | 64 | 0.001597 | N/A | ReLU | 100 |
| | LSTM | 300 | 200 | N/A | N/A | 16 | 0.3074 | 0.3568 | Tanh | 50 |
| *BPE* | MLP | 64 | 300 | N/A | N/A | 64 | 0.00113 | 0.3273 | Tanh | 200 |
| | CNN | 200 | N/A | 3, 4, 5 | 256 | 64 | 0.00009431 | N/A | ReLU | 100 |
| | LSTM | 200 | 200 | N/A | N/A | 32 | 0.4259 | 0.2247 | Tanh | 200 |
| *LIWC* | MLP | 300 | 100 | N/A | N/A | 64 | 0.03566 | 0.2758 | ReLU | 200 |
| | CNN | 100 | N/A | 3, 4, 5 | 128 | 64 | 0.0187 | N/A | Tanh | 100 |
| | LSTM | 300 | 200 | N/A | N/A | 16 | 0.05504 | 0.224 | Tanh | 50 |

**Table 5.14** Best hyper-parameters for models optimised on the *Toxicity* dataset.

| | Model | Embedding Dim | Hidden Dim | Window Size | Filters | Batch Size | Learning Rate | Dropout | Non-linearity | # Epochs |
|---|---|---|---|---|---|---|---|---|---|---|
| *WORD* | MLP | 100 | 64 | N/A | N/A | 16 | 0.3896 | 0.1711 | Tanh | 50 |
| | CNN | 200 | N/A | 1, 2, 3 | 256 | 32 | 0.002737 | N/A | Tanh | 200 |
| | LSTM | 300 | 300 | N/A | N/A | 16 | 0.4796 | 0.4006 | Tanh | 50 |
| *BPE* | MLP | 64 | 200 | N/A | N/A | 16 | 0.7323 | 0.3823 | Tanh | 200 |
| | CNN | 300 | N/A | 2, 3, 4 | 256 | 64 | 0.01062 | N/A | ReLU | 200 |
| | LSTM | 64 | 64 | N/A | N/A | 32 | 0.9958 | 0.1481 | Tanh | 200 |
| *LIWC* | MLP | 200 | 100 | N/A | N/A | 64 | 0.1487 | 0.0901 | Tanh | 50 |
| | CNN | 300 | N/A | 2, 3, 4 | 64 | 16 | 0.01134 | N/A | Tanh | 100 |
| | LSTM | 200 | 300 | N/A | N/A | 32 | 0.7015 | 0.2012 | Tanh | 100 |

**Table 5.15** Best hyper-parameters for models optimised on the *Hate Expert* dataset.

| | Model | Embedding Dim | Hidden Dim | Window Size | Filters | Batch Size | Learning Rate | Dropout | Non-linearity | # Epochs |
|---|---|---|---|---|---|---|---|---|---|---|
| *WORD* | MLP | 300 | 200 | N/A | N/A | 64 | 0.5251 | 0.2284 | Tanh | 200 |
| | CNN | 200 | N/A | 1, 2, 3 | 64 | 16 | 0.01644 | N/A | Tanh | 50 |
| | LSTM | 64 | 300 | N/A | N/A | 32 | 0.6271 | 0.02672 | Tanh | 50 |
| *BPE* | MLP | 200 | 300 | N/A | N/A | 64 | 0.0005028 | 0.3471 | ReLU | 50 |
| | CNN | 200 | N/A | 1, 2, 3 | 128 | 64 | 0.002686 | N/A | ReLU | 100 |
| | LSTM | 300 | 300 | N/A | N/A | 16 | 0.2446 | 0.03933 | Tanh | 200 |
| *LIWC* | MLP | 300 | 100 | N/A | N/A | 32 | 0.07654 | 0.399 | Tanh | 200 |
| | CNN | 64 | N/A | 2, 3, 4 | 64 | 32 | 0.001567 | N/A | ReLU | 200 |
| | LSTM | 200 | 64 | N/A | N/A | 32 | 0.04429 | 0.305 | Tanh | 200 |

**Table 5.16** Best hyper-parameters for models optimised on the *Hate Speech* dataset.

| | Model | Embedding Dim | Hidden Dim | Window Size | Filters | Batch Size | Learning Rate | Dropout | Non-linearity | # Epochs |
|---|---|---|---|---|---|---|---|---|---|---|
| *WORD* | MLP | 64 | 200 | N/A | N/A | 32 | 0.3926 | 0.438 | Tanh | 200 |
| | CNN | 200 | N/A | 1, 2, 3 | 64 | 16 | 0.1614 | N/A | ReLU | 200 |
| | LSTM | 200 | 100 | N/A | N/A | 32 | 0.8455 | 0.3605 | Tanh | 100 |
| *BPE* | MLP | 64 | 100 | N/A | N/A | 32 | 0.000325 | 0.1971 | ReLU | 200 |
| | CNN | 200 | N/A | 3, 4, 5 | 256 | 64 | 0.009577 | N/A | ReLU | 100 |
| | LSTM | 100 | 300 | N/A | N/A | 16 | 0.5939 | 0.4588 | Tanh | 200 |
| *LIWC* | MLP | 100 | 100 | N/A | N/A | 64 | 0.03136 | 0.2871 | Tanh | 200 |
| | CNN | 200 | N/A | 3, 4, 5 | 128 | 16 | 0.001456 | N/A | Tanh | 200 |
| | LSTM | 100 | 300 | N/A | N/A | 32 | 0.05931 | 0.1301 | Tanh | 200 |

**Table 5.17** Best hyper-parameters for models optimised on the *StormFront* dataset.

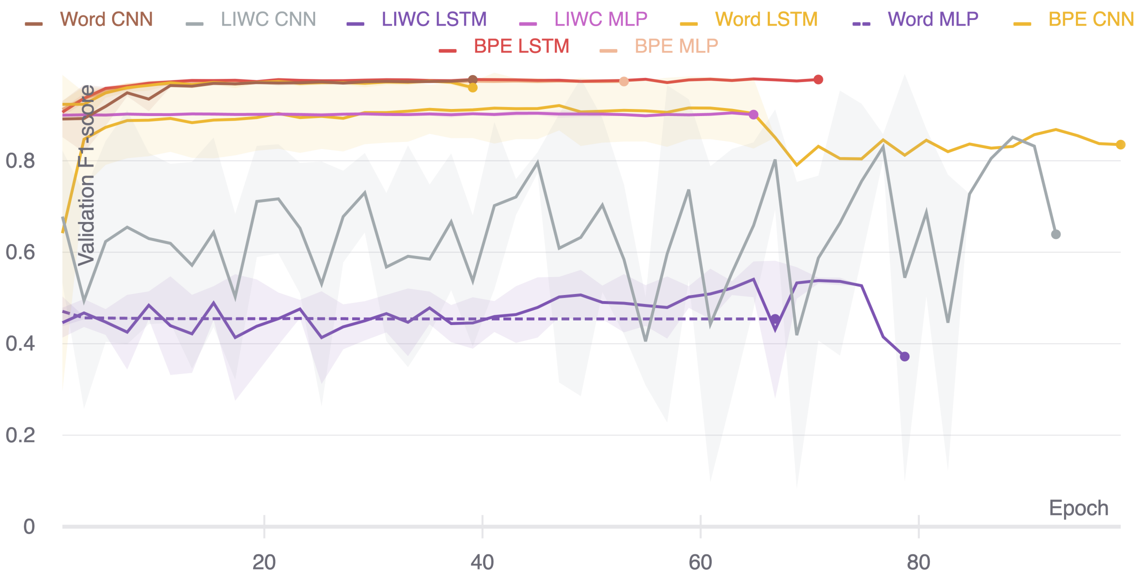


**Fig. 5.1** In-domain macro `F1-score` on validation set for models optimised on the *Offence* dataset.

In observing the `F1-scores` on the validation sets in figs. 5.1, 5.3, 5.5, 5.7 and 5.9, a general pattern emerges in which most models, regardless of input type, display similar learning curves. There are however some notable exceptions to this rule. One such exception can be observed in fig. 5.1, where the LIWC-based CNN displays volatile performances throughout the entire optimisation procedure. Another exception can be observed in fig. 5.5, where the BPE-based MLP model starts with a low performance, but increasingly improves until the final few epochs, where the model predictions become very volatile and ultimately triggers early stopping with a large drop in performance. In addition to these two exceptions, the models that are optimised for the *StormFront* dataset (see fig. 5.9) display three patterns: First, there are some models that show a large variability in their performances from epoch to epoch, these models tend to trigger early stopping at an early stage. Second, there are models that show a smaller degree of variability in classification performance as the model is optimised, but as they pass through the epochs, the model performance steadily increases and the variability in performances decreases. Third, there are models that obtain a high score early in the optimisation procedure and trigger early stopping at an early stage.

Another exception can be observed in the models optimised on *Toxicity* dataset (see fig. 5.3). Here three salient trends occur. In the first, models start with a high `F1-score` and show little improvement as over the epochs and trigger early stopping. In the second, models start with a lower `F1-score` and show steady improvements until early stopping is triggered. The third trend starts with a relatively low model performance (below 0.4 in `F1-score`), seeing steady

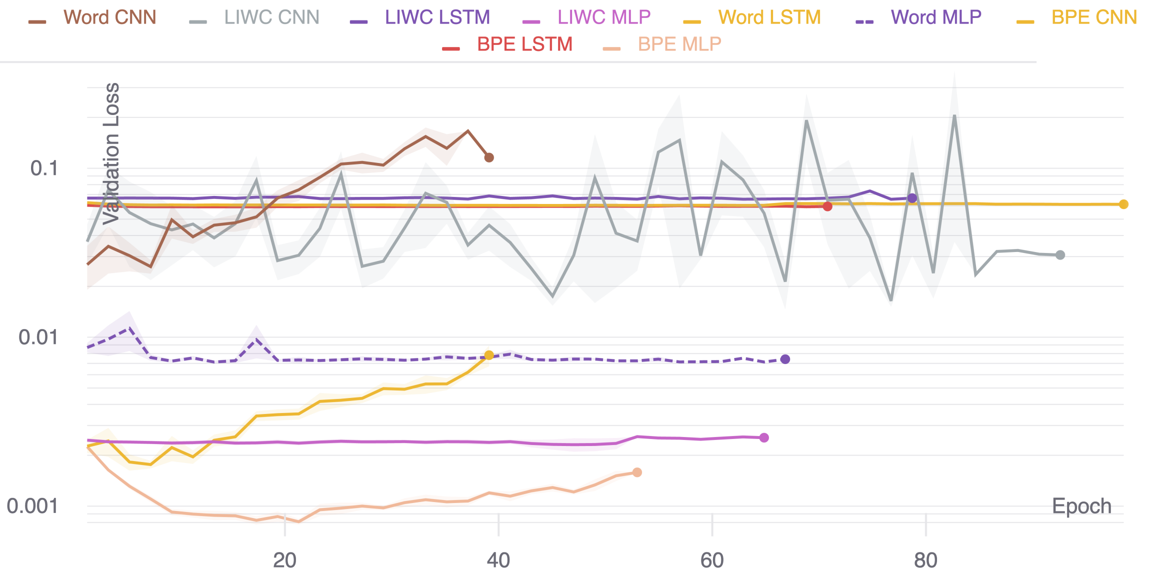


**Fig. 5.2** Validation losses for models optimised on the *Offence* dataset.

improvements, before reaching a plateau and continuing until early stopping is triggered or the set number of epochs is reached.

Turning to the loss developments in figs. 5.2, 5.4, 5.6, 5.8 and 5.10 there are three unique patterns: First, the loss rises throughout the optimisation process. Second, the loss remains almost entirely unchanged throughout the entire optimisation process. Finally the third pattern, the loss is volatile throughout the process, rising or dropping from epoch to epoch.

#### 5.3.3.1 Evaluation Set Performances

In figs. 5.11 to 5.13, 5.15 to 5.18, 5.20 to 5.23 and 5.25, I show the in-domain and out-of-domain results of using the neural network architectures described in section 5.2 for modelling abuse using the three different document representations. The bars in each figure represent the scores of each models on the test set in question, e.g. in fig. 5.11 I show the macro `F1-scores` achieved by all models on the test set for the *Offence* dataset, and the error bars are the standard deviation over the 5 parameter seed runs.

Considering figures collectively, it's clear that in-domain models in most cases, predictably, out-perform models optimised on out-of-domain datasets. Interestingly, it's also clear that LIWC-based models in many cases are comparable to models using full surface-form vocabulary. Moreover, this similarity in performance largely also holds for out-of-domain performance, with some LIWC-based models consistently ranking among the best performing out-of-domain models. The comparison of out-of-domain performance between experimental

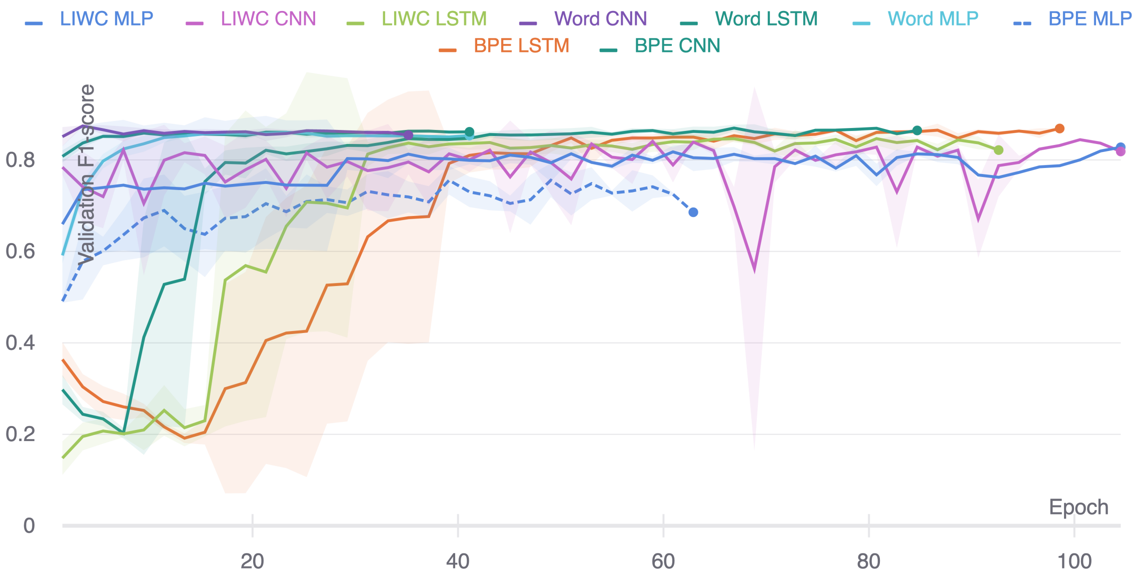


**Fig. 5.3** In-domain macro `F1-score` on validation set for models optimised on the *Toxicity* dataset.

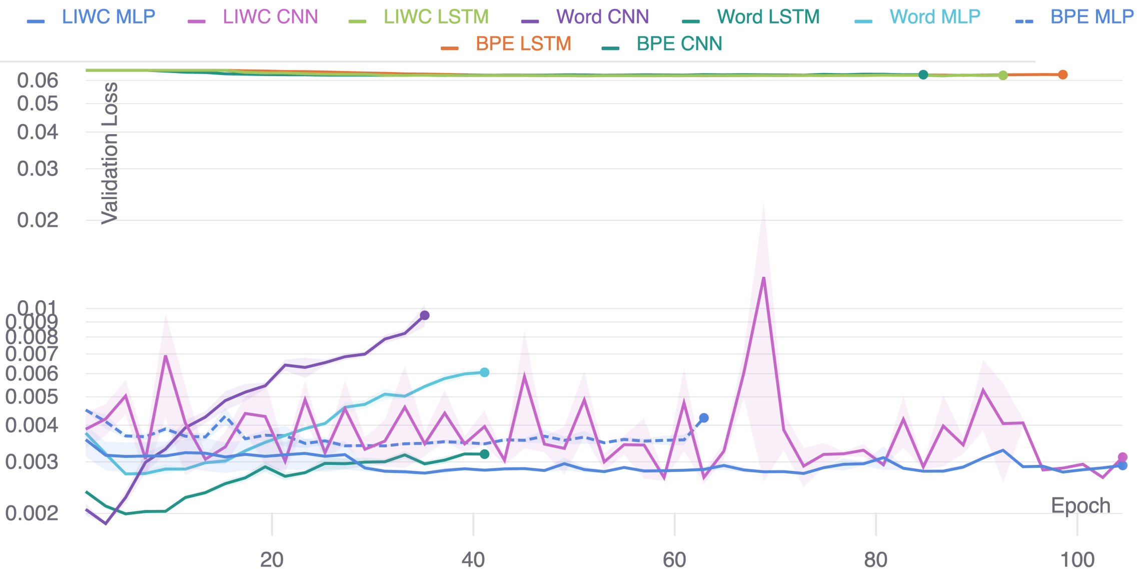


**Fig. 5.4** Validation losses for models optimised on the *Toxicity* dataset.

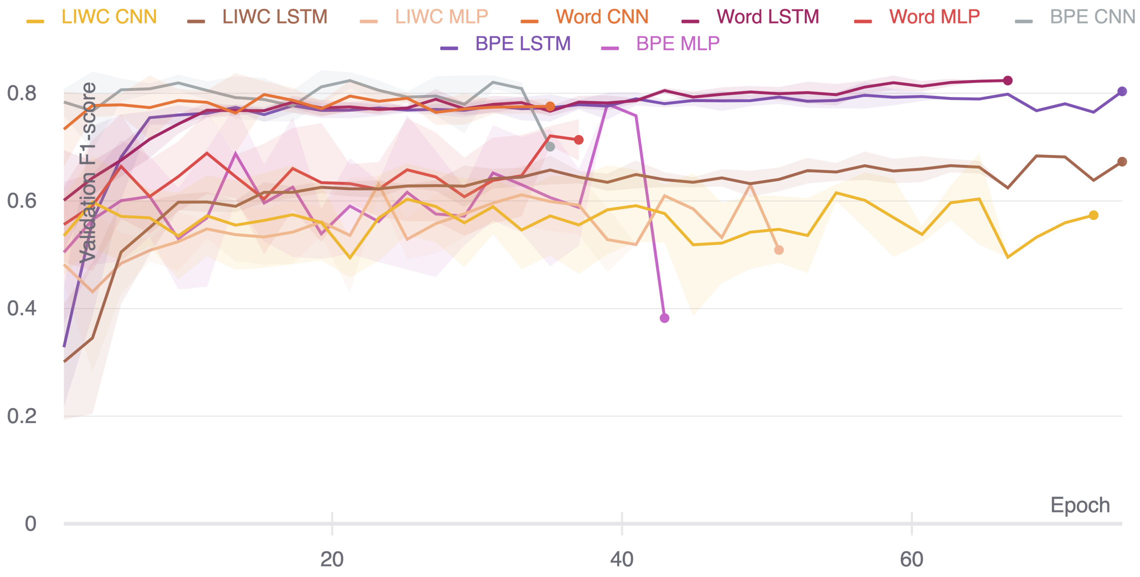


**Fig. 5.5** In-domain macro `F1-score` on validation set for models optimised on the *Hate Expert* dataset.

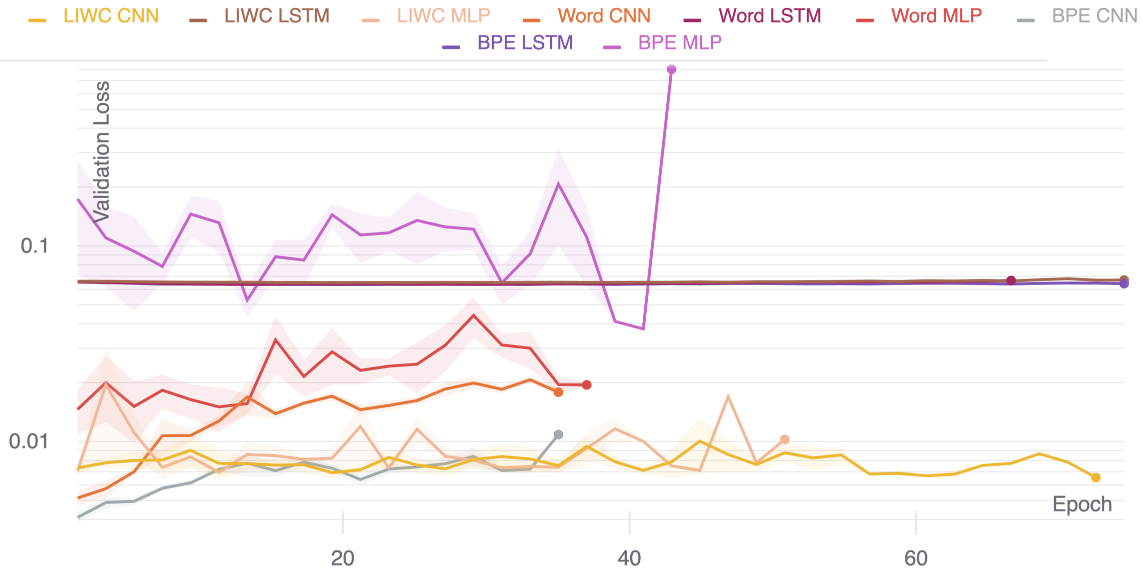


**Fig. 5.6** Validation losses for models optimised on the *Hate Expert* dataset.

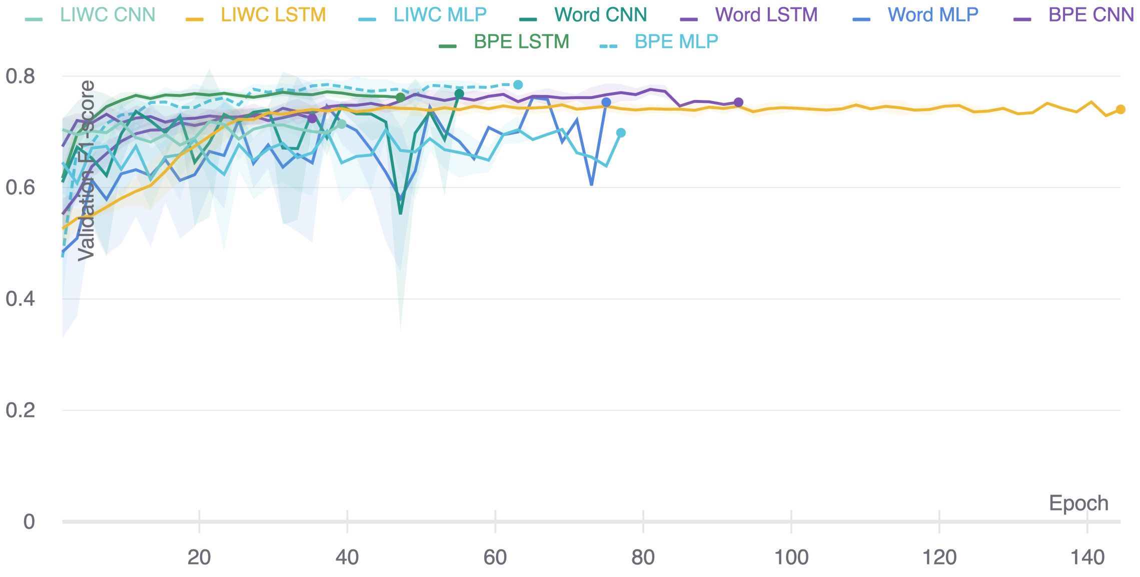


**Fig. 5.7** In-domain macro `F1-score` on validation set for models optimised on the *Hate Speech* dataset.

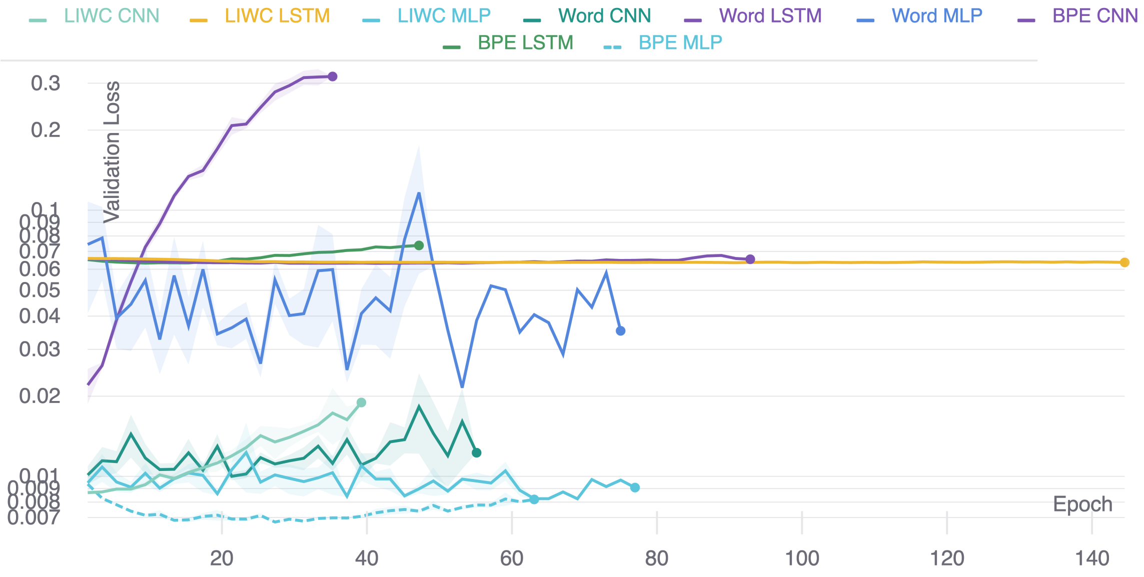


**Fig. 5.8** Validation losses for models optimised on the *Hate Speech* dataset.

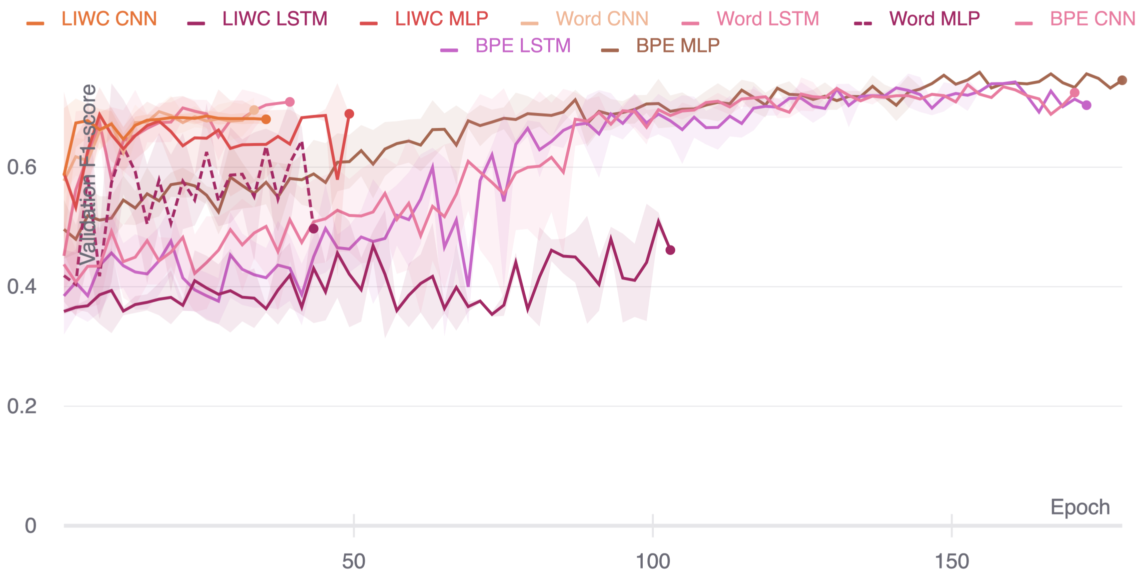


**Fig. 5.9** In-domain macro `F1-score` on validation set for models optimised on the *StormFront* dataset.

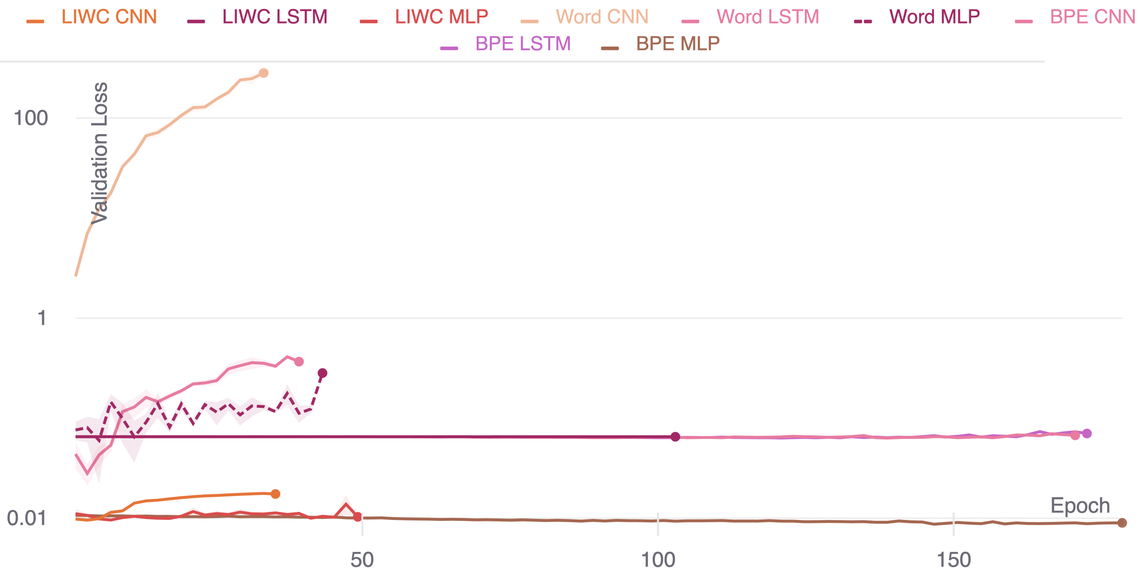


**Fig. 5.10** Validation losses for models optimised on the *StormFront* dataset.

models and the linear baselines too is worth noting. Here, in spite of improved performances on the in-domain evaluation sets, there is a tend towards a slight decrease performance on out-of-domain data by the experimental models.

One dataset however, is notable in its in-domain and all out-of-domain predictions: The *StormFront* dataset. For this dataset, linear models perform at par, or better than all configurations of neural models. The most likely explanation for this can be found in the small dataset size of less than $3,000$ documents. One way to address such a short-coming of this dataset is to increase the dataset size. Although the experiments I conduct with the dataset keep the number of documents lower than the total annotated set in order to maintain a balanced dataset, the dataset does in some cases allow for models that out-perform models optimised on larger datasets, in terms of out-of-domain prediction. Specifically, the MLP models evaluated on the *Offence* data (see fig. 5.11, where the StormFront LIWC-based model out-performs several other models that are optimised on larger datasets.

More generally, from the out-of-domain classification performances, there seems to be a correlation with the goals of the datasets and out-of-domain performance. For instance, in figs. 5.11, 5.12, 5.16, 5.21 and 5.22, I observe that models optimised on the *Offence* and *Toxicity* datasets out-perform models optimised on other datasets. For both of these datasets, the governing understanding of abuse and hate speech are that not all speech that is offensive is necessarily also problematic. The motivation for the development of the *Offence* dataset was specifically to disentangle hateful from offensive but not necessarily unacceptable. Similarly, the *Toxicity* dataset asked its annotators to identify comments that might make people exit conversations they were part of rather than ask annotators to label for all content that is offensive or hateful. Thus, for these two datasets, the governing question is not necessarily the protection of marginalised communities and identities but instead identifying a degree of acceptable abuse and hostility.

In contrast, the *Hate Expert* and *Hate Speech* datasets seek to identify communications that are harmful to marginalised communities. Thus, it's no surprise that the out-of-domain performances for models optimised on these two datasets perform reasonably well with each other. However here it is also clear that there is a relationship between dataset size and out-of-domain performance. The models optimised on the larger dataset (i.e. *Hate Speech* with $16,000$ documents) have better performance on the evaluation set of the smaller (the *Hate Expert* dataset with $7,000$ documents) than the other way around.

The *StormFront* dataset on the other hand is annotated to identify deliberate attacks against "specific group[s] of people" on the basis of their group membership or characteristics of group's identities (Garcia et al., 2019). This annotation criteria forms a subset of the annotation guidelines that are used for the *Hate Speech* and *Hate Expert* datasets. Moreover, the collection strategies for the three datasets also share common characteristics. Where Garcia et al. (2019) specifically seek out content from a white supremacist forum for their dataset, Talat (2016); Talat and Hovy (2016) sample from Twitter by searching for keywords that were likely to result in a large set of of gendered and racialised abuse. Thus, while the domain of the data and the annotation guidelines are not the same, there are likely to be similarities in the content and annotations produced.

Attending to the questions surrounding the use of LIWC to represent documents for modelling abuse, I turn to the baseline and experimental model performances on the validation and evaluation sets of the LIWC-based models. In the validation sets for the linear baselines (see table 5.7), the LIWC-based methods do not out-perform any other model type on some datasets while it does on others. For instance, a highly competitive score is obtained (e.g. 0.9207 for LIWC-based SVM model against 0.9222 for a word token-based SVM) . For other datasets however, the score obtained by LIWC-based models is much lower, suggesting that LIWC-based modelling may be an appropriate means of modelling abuse under some conditions. For the neural network-based models a similar story presents itself, although the LIWC-based models perform reasonably well in comparison to models that use a larger vocabulary (e.g. 0.9644 for the LIWC-based model in table 5.13 and 0.9783 for the BPE-based model).

On the evaluation sets however, a slightly different patterns plays out. For most evaluation datasets, at least one of the LIWC-based models out-perform some of the surface token-based models, and in some cases out-perform all other models. In particular, fig. 5.12, the in-domain LIWC-based model out-performs all other model types, though notably posts lower scores than the out-of-domain LIWC-based *Offence* model.

Overall, the patterns displayed by the MLPs (see figs. 5.11 to 5.15) indicate that LIWC-based document representations are appropriate for the development of neural networks for abuse detection, in spite of the large reduction in vocabulary size.

In fact, the model performances also weakly indicate that there are benefits to be found in using LIWC-based representations for medium sized and large datasets.[3] Considering the LIWC-based *Toxicity* and *Hate Speech* models, their performance appears slightly less volatile to domain shifts in comparison to their surface token counterparts, although they still exhibit a high degree of volatility as the goals of the datasets change.

[3] Medium and large are relative here to to the sizes of hate speech and abuse data, often ranging between $5,000$ to $100,000$ samples, rather than large scale datasets for computing that contain millions of samples.

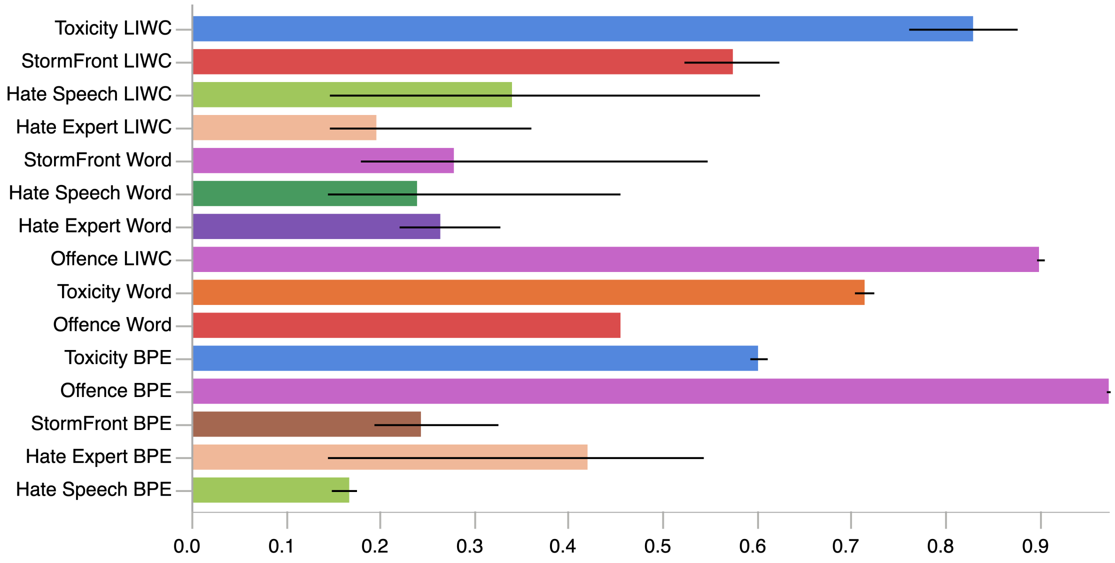


**Fig. 5.11** Macro `F1-scores` for all MLP models on the *Offence* evaluation set with the standard deviation represented in error bars.

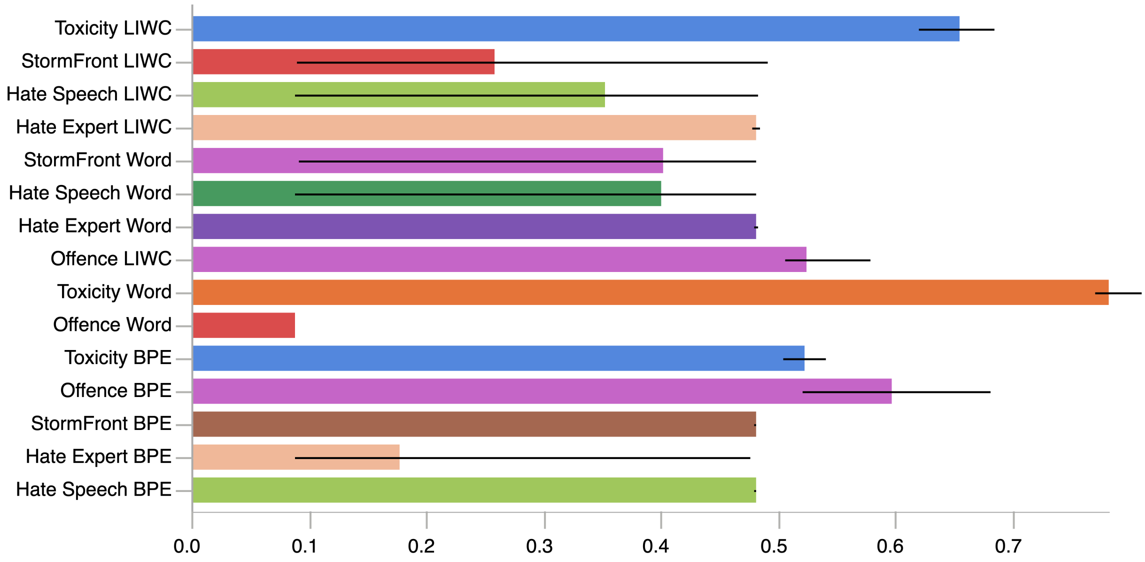


**Fig. 5.12** Macro `F1-scores` for all MLP models on the *Toxicity* evaluation set with the standard deviation represented in error bars.

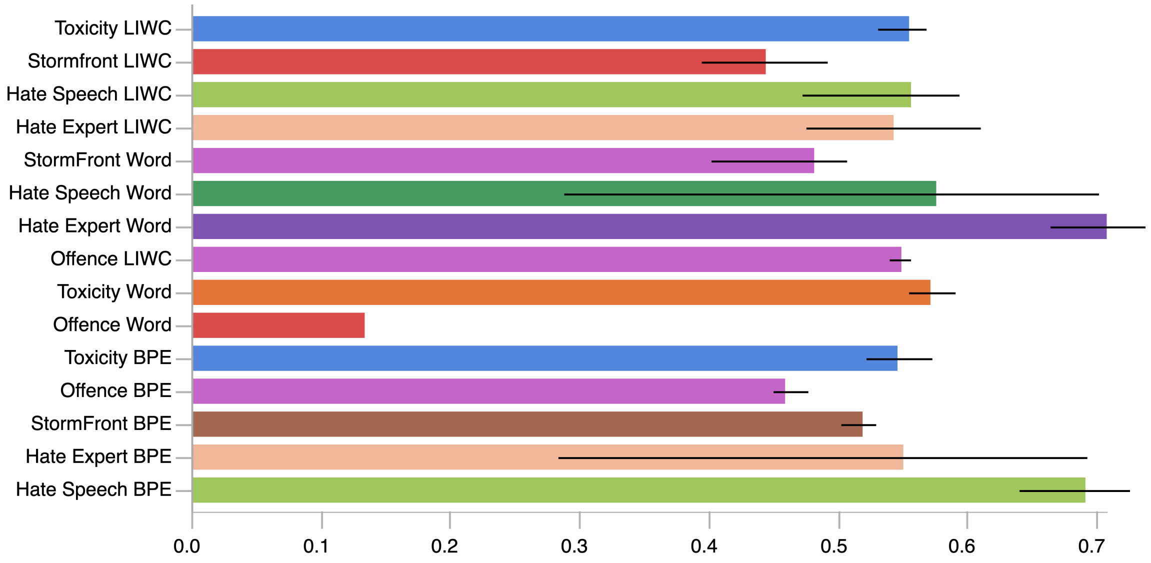


**Fig. 5.13** Macro `F1-scores` for all MLP models on the *Hate Expert* evaluation set with the standard deviation represented in error bars.

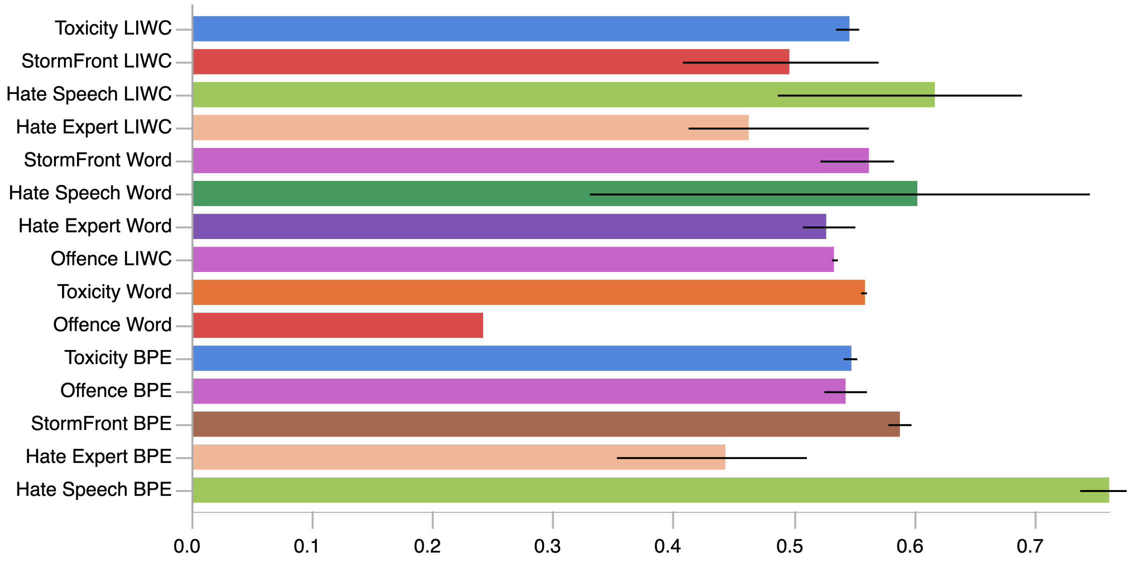


**Fig. 5.14** Macro `F1-scores` for all MLP models on the *Hate Speech* evaluation set with the standard deviation represented in error bars.

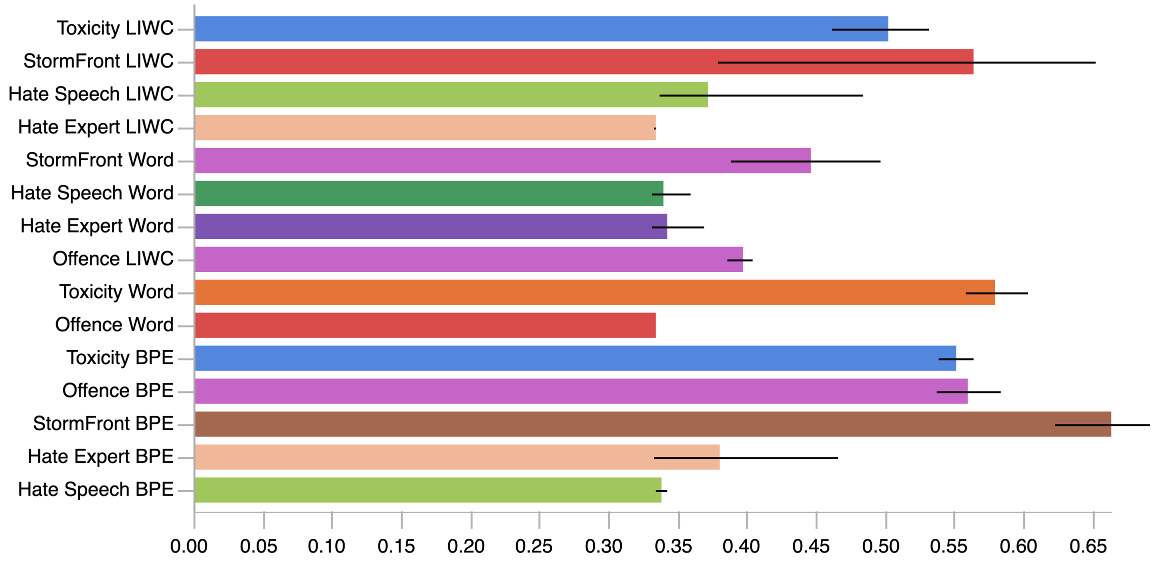


**Fig. 5.15** Macro `F1-scores` for all MLP models on the *StormFront* evaluation set with the standard deviation represented in error bars.

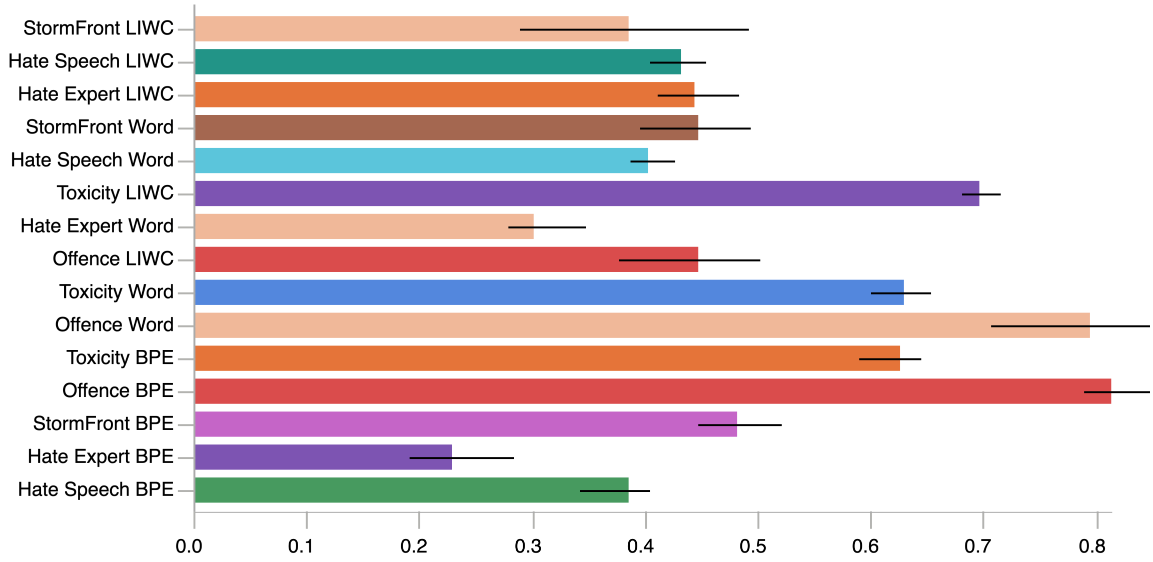


**Fig. 5.16** Macro `F1-scores` for all LSTM models on the *Offence* evaluation set with the standard deviation represented in error bars.

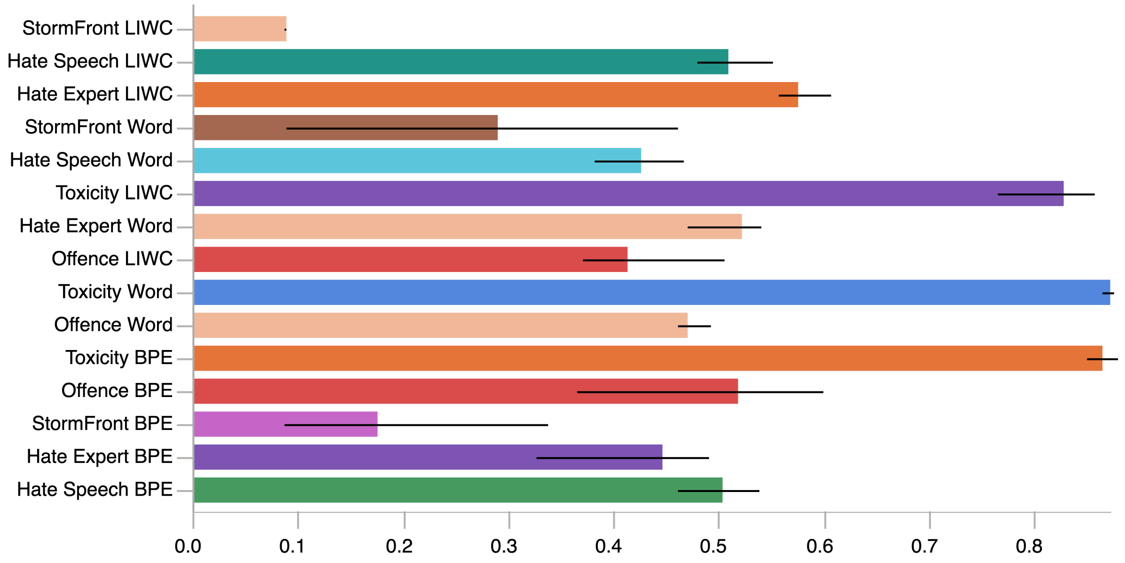


**Fig. 5.17** Macro `F1-scores` for all LSTM models on the *Toxicity* evaluation set with the standard deviation represented in error bars.

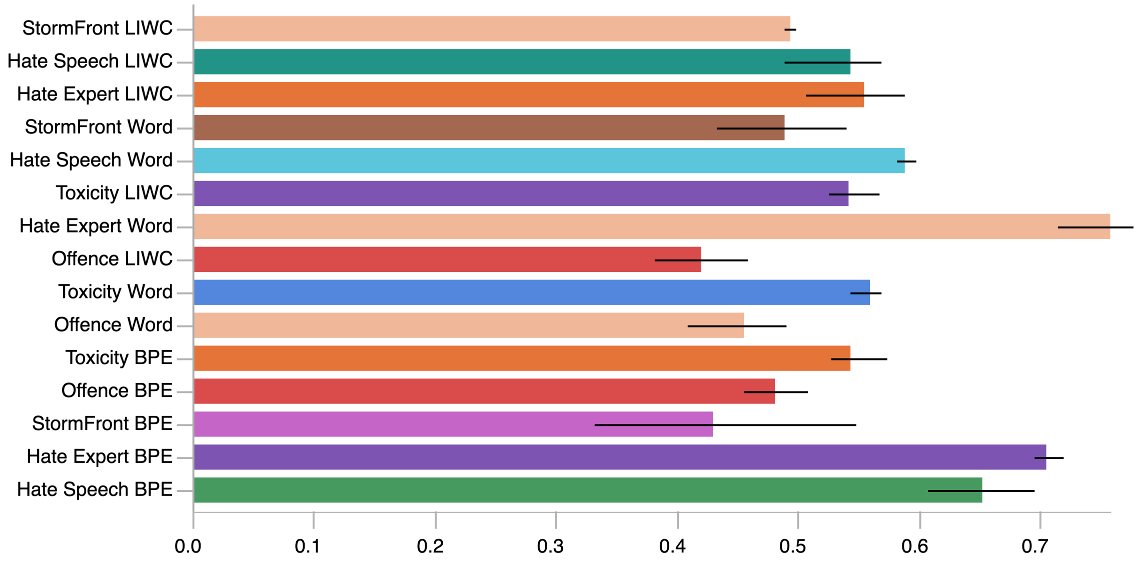


**Fig. 5.18** Macro `F1-scores` for all LSTM models on the *Hate Expert* evaluation set with the standard deviation represented in error bars.

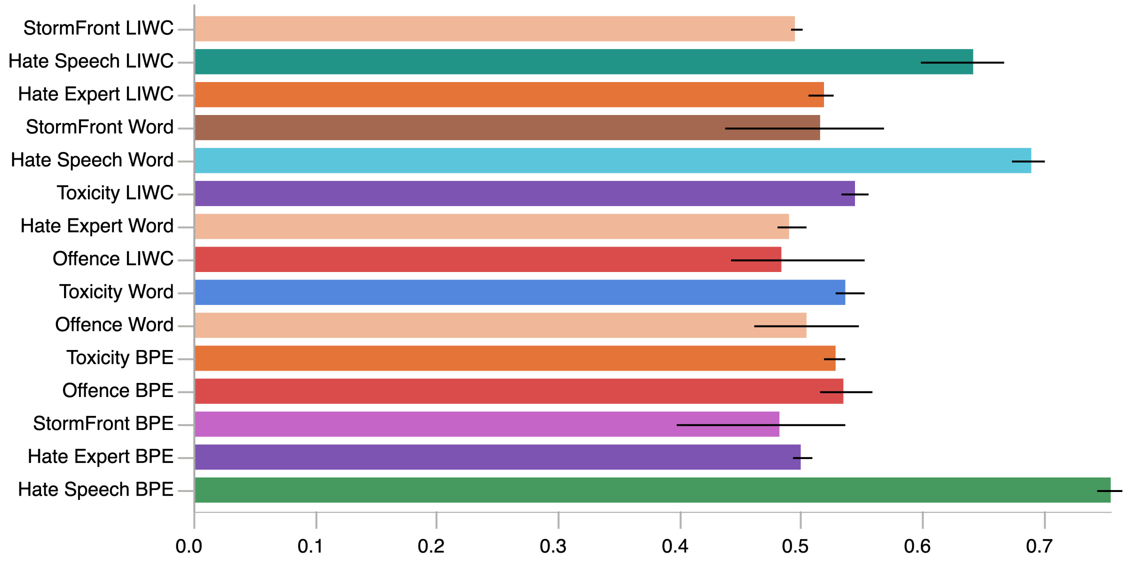


**Fig. 5.19** Macro `F1-scores` for all LSTM models on the *Hate Speech* evaluation set with the standard deviation represented in error bars.

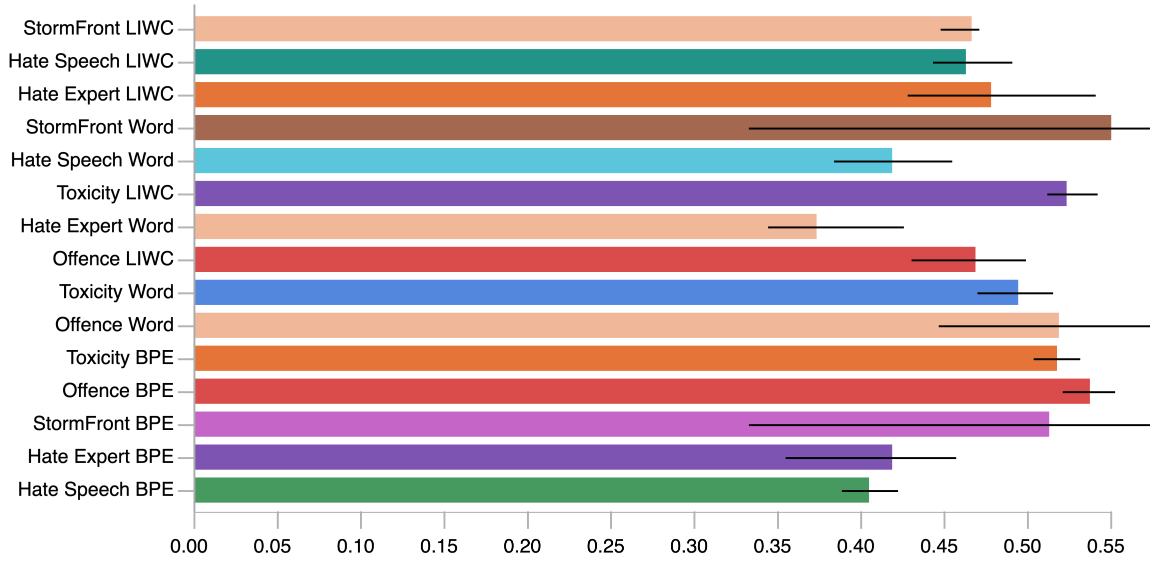


**Fig. 5.20** Macro `F1-scores` for all LSTM models on the *StormFront* evaluation set with the standard deviation represented in error bars.

As I can establish that LIWC-based data representation is appropriate for neural network methods, I turn to ask what the influence of recurrence is on the performance of models when predicting on in-domain and out-of-domain datasets through the use of LSTM models.

On examining the test scores on the LSTM models (see figs. 5.16 to 5.20), I find that the performance of MLPs and LSTMs in general are competitive with one another and it is dependent on the dataset, which architecture will work best. To the point of this chapter, while there is some variability in the performance of LIWC-based models in some instances, they achieve high in-domain and out-of-domain performances. This for instance is the case with the LIWC-based model optimised on the *Toxicity* dataset, which consistently achieves a high *F1-score* on the *Offence* dataset.

Comparing only the LIWC-based models with each other, I find that in general, MLPs are out-performed by LSTM models on in-domain data. The LSTM models, in turn tend to achieve lower performances than the CNN models. In most cases, all models obtain high performances and are competitive with one another. Although it is slightly surprising that recurrence seems to have only have a small positive effect on the in-domain performances of the LSTM models, this follows the prior work in abuse detection where CNN models long had a dominance over other models due to their ability to outperform most other models.

Taking into consideration the generally high performance of the LIWC-based models optimised for the *Toxicity* dataset, there appears to be an effect between the size of the dataset and the in-domain and out-of-domain performances. The LIWC-based models optimised on the *Toxicity* dataset tend to out-perform other models optimised on the *Toxicity* dataset, when evaluated on out-of-domain datasets. These results suggest that LIWC-based modelling may provide for improved out-of-domain performances when the models are optimised on large datasets or are being applied to data with which there are shared attributes in terms of annotation goal.

Comparing the LIWC-based models on out-of-domain performance, I note that for LIWC-based models similarities in dataset goals seems to have a positive effect. Additionally, the models optimised on the *Offence* tend to perform well across the different datasets.

#### 5.3.3.2 Computational costs

Reducing the size of the vocabulary may also have implications for time required to optimised models, which can have downstream effects on the environmental impacts of developing machine learning models for abuse detection. Here, I consider the impacts that using LIWC-based document representations have on optimisation time. Figures 5.26 to 5.30 show the

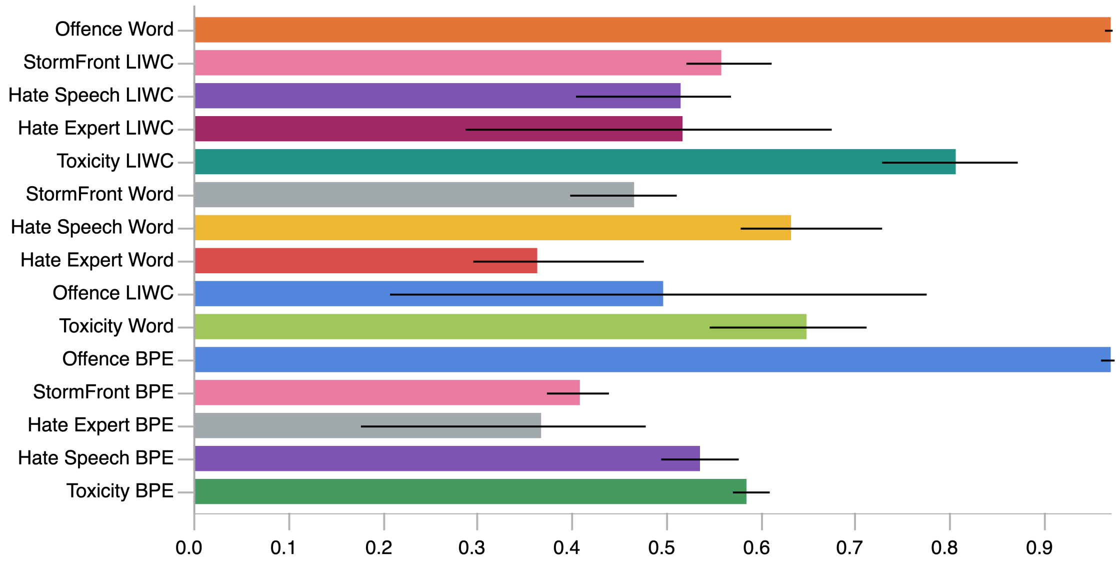


**Fig. 5.21** Macro `F1-scores` for all CNN models on the *Offence* evaluation set with the standard deviation represented in error bars.

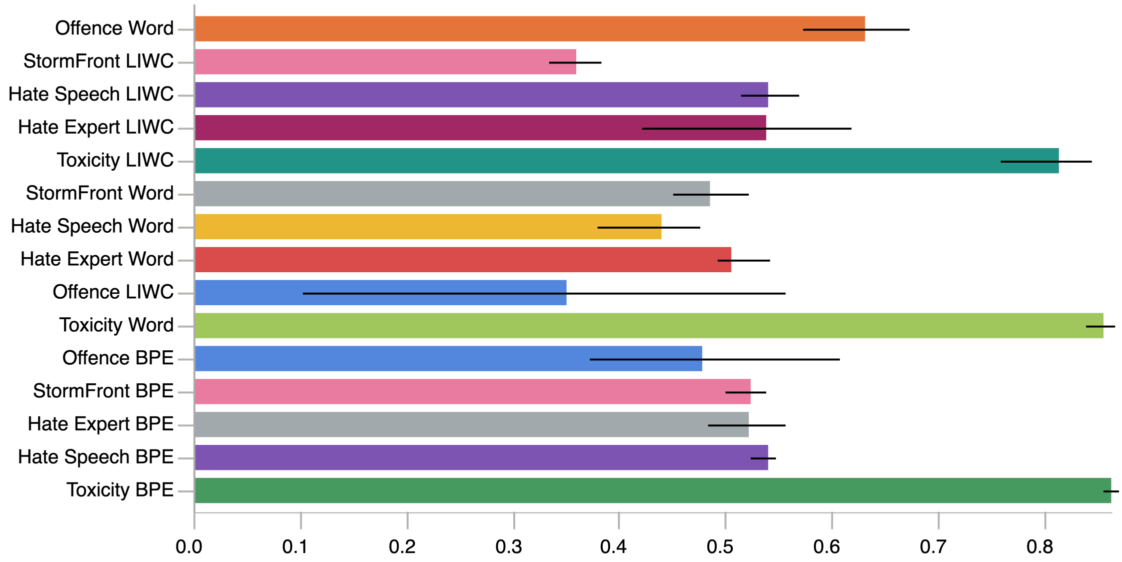


**Fig. 5.22** Macro `F1-scores` for all CNN models on the *Toxicity* evaluation set with the standard deviation represented in error bars.

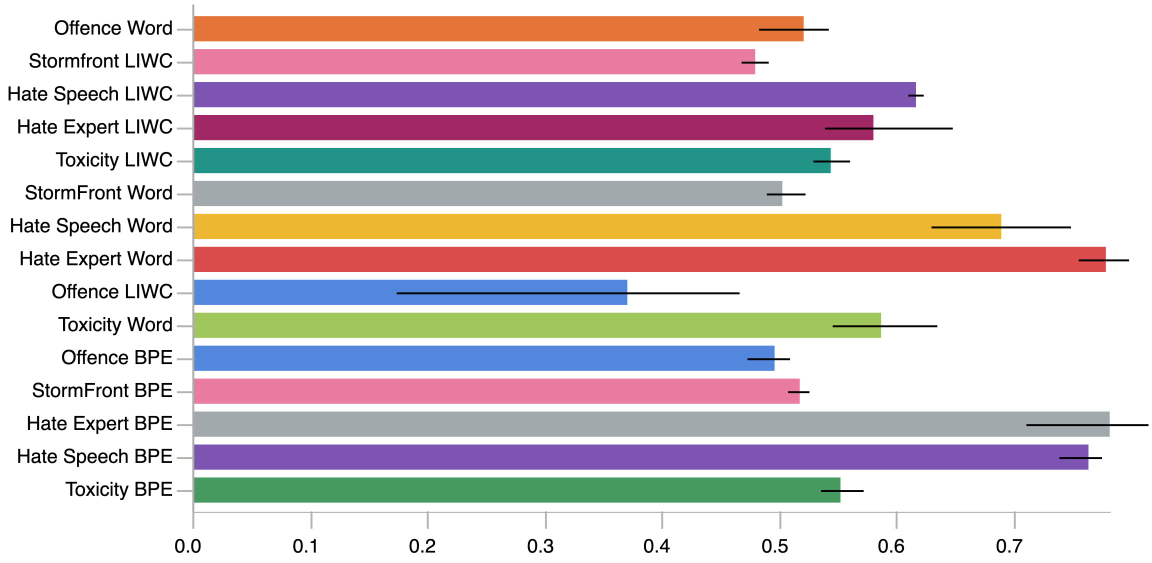


**Fig. 5.23** Macro `F1-scores` for all CNN models on the *Hate Expert* evaluation set with the standard deviation represented in error bars.

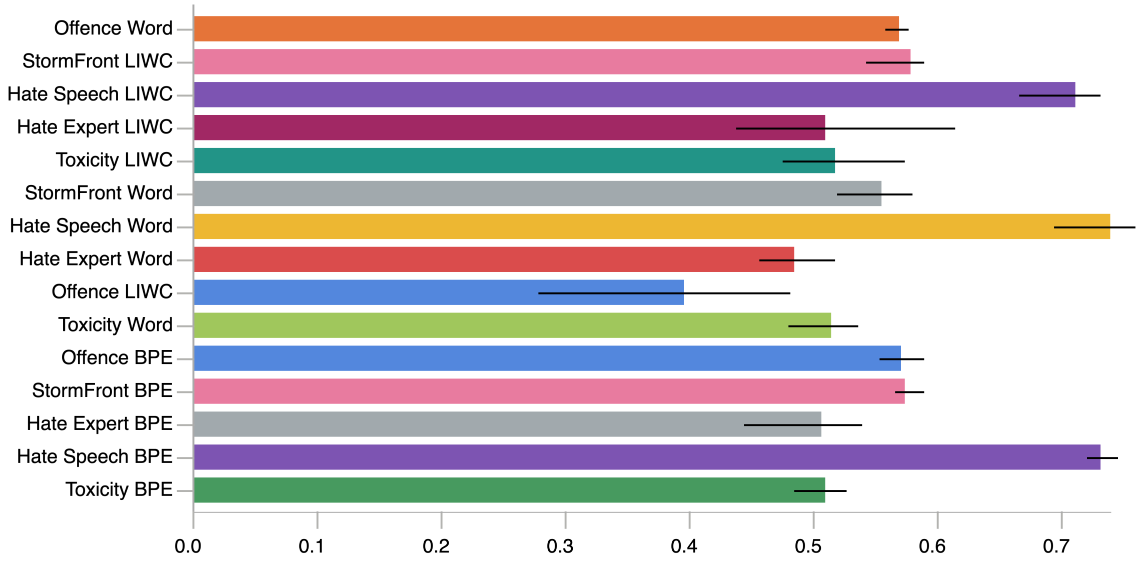


**Fig. 5.24** Macro `F1-scores` for all CNN models on the *Hate Speech* evaluation set with the standard deviation represented in error bars.

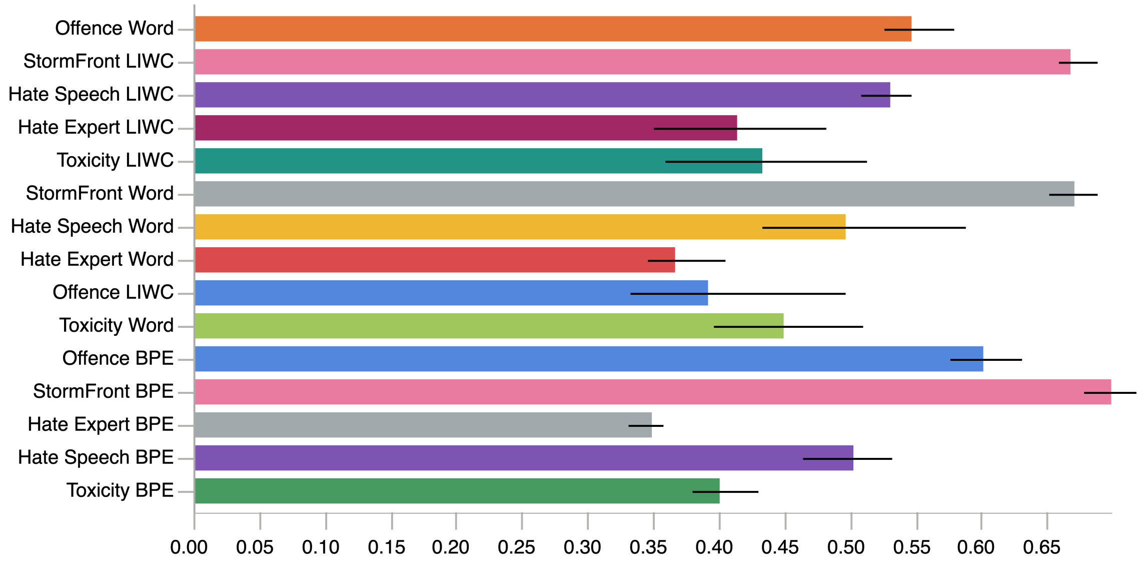


**Fig. 5.25** Macro `F1-scores` for all CNN models on the *StormFront* evaluation set with the standard deviation represented in error bars.

number of minutes taken for each model to optimised on each dataset with the error bars representing the standard deviation across 5 runs.

First, there is a predictable correlation with the complexity of the machine learning model and the time required to optimise a model, with the MLP models being the quickest to be optimised and the LSTM models taking the longest, with the exception of the CNN models optimised on the *Hate Expert* dataset, where the LIWC CNN requires roughly twice as long as the LIWC LSTM to optimise (see fig. 5.28).

Considering the influence of document representation on optimisation time, the results point in multiple directions. First, figs. 5.26 to 5.30 show that LIWC-based representation for MLPs and CNNs, in most cases, yields faster optimisation time than when using the surface forms. The figures also show that LSTMs that use LIWC tend to finish optimising faster than the LSTMs optimised on the surface forms. However, on the largest dataset, the *Toxicity* dataset, the LIWC-based LSTM is slower to finish optimising than it's surface form counter parts. The LIWC-based CNN is slower to optimise than the word token CNN but faster than the BPE CNN. Finally, the LIWC-based MLP optimises slightly faster than the models for word token input and BPE input.

In the medium sized datasets however, the relation between optimisation time and vocabulary minimisation is clear as the LIWC-based models tend to take less time to optimise than counter-part, as is apparent  in figs. 5.26, 5.28 and 5.29. For each of these datasets, some

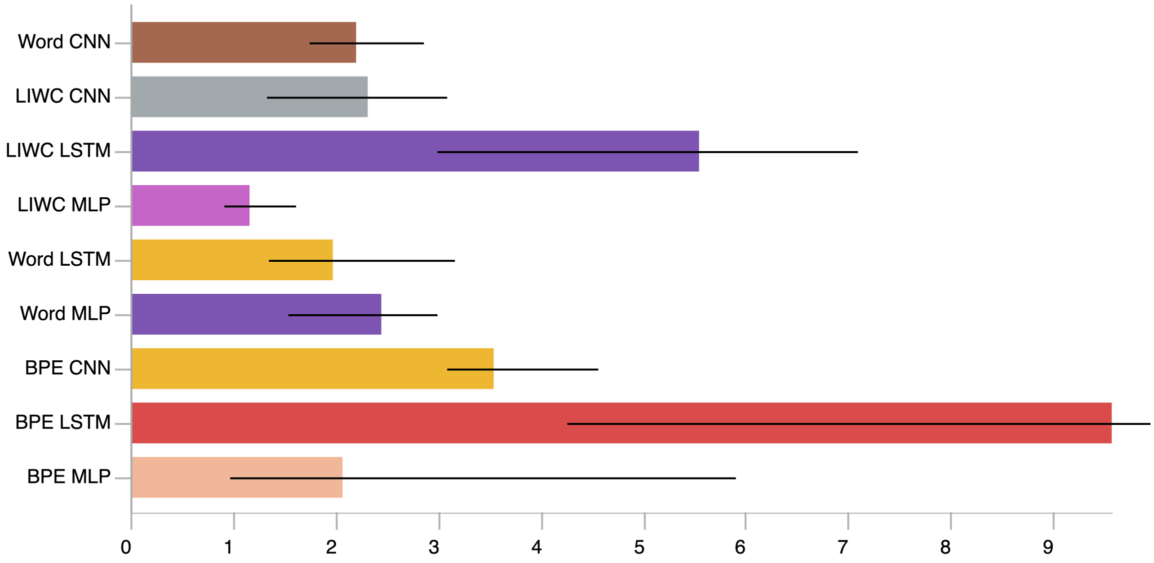


**Fig. 5.26** Optimisation time in minutes for each model type on the *Offence* dataset.

LIWC-based models are optimised quicker while others are slower. In part, this appears to be connected to model complexity, where the more complex the underlying model is, the slower the optimisation time also is. For instance, in fig. 5.28 the LIWC-based MLP is the quickest MLP to be optimised while the LIWC-based LSTM models is slower than all other LSTM models.

Reflecting on the feasibility of using LIWC-based document representations for optimising neural networks, I turn to the two largest datasets, the *Offence* and *Toxicity* datasets, and the performance of LIWC-based models while bearing in mind their optimisation time (see table 5.18). Beyond being the two largest datasets, I choose these to compare as their operationalisations of abuse share large similarities. Thus, one can consider each dataset a domain shifted dataset to the other, thus providing a more reasonable point of comparison than using a dataset which is annotated with a fundamentally different goal.

In table 5.18, it immediately stands out that some LIWC-based models take longer to optimise than their surface token-based counter-parts. Moreover, It is clear that many of the surface token-based models that perform very well on the in-domain evaluation set do not see the performance transfer to other datasets. Conversely, some LIWC-based models see a lesser drop in performance on external data, if not an outright increase in performances. This suggests that LIWC-based modelling may capture more general patterns of abuse that models are otherwise prone to over-fit away from in pursuit of improved in-domain performance.

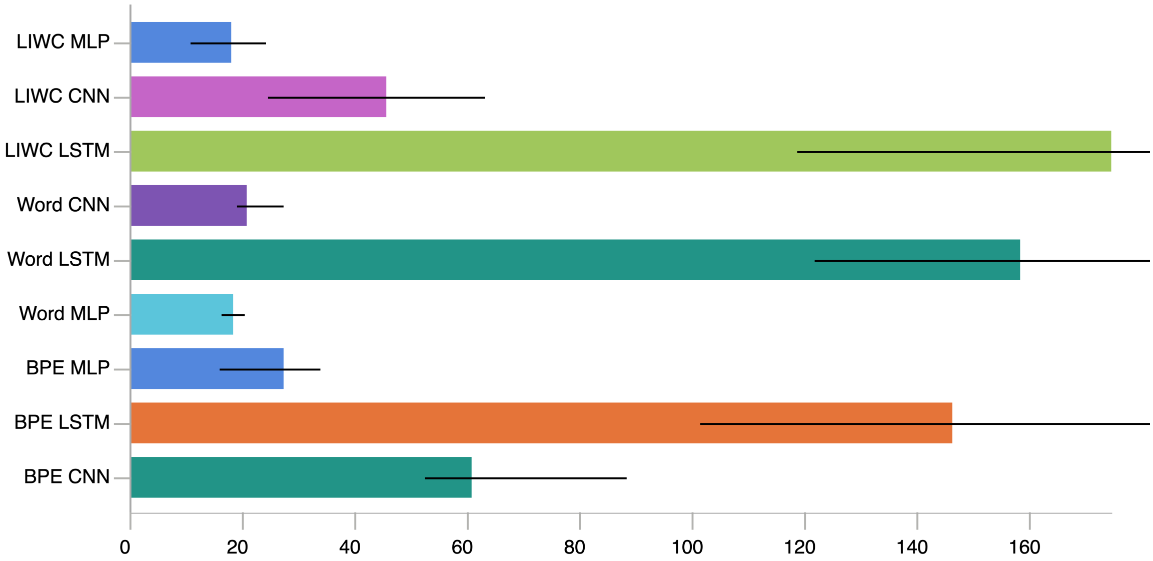


**Fig. 5.27** Optimisation time in minutes for each model type on the *Toxicity* dataset.

### 5.3.4 A Consideration through Dirt

I return here to the considerations in chapter 4 to understand how LIWC-based representation comes to influence the challenges identified in chapter 4. Similarly to the Perspective API, the LIWC-based model take a top-down approach to determining what constitutes ‘abuse’. For this reason, a great deal of the analysis provided in section 4.2.1 also holds here. Specifically, the use of similar neural network-based models are based on similar foundations, where my models deviate is through the use of LIWC Pennebaker et al. (2001) to transform the input data to the LIWC categories that are invoked, where the Perspective API uses surface level representations of tokens without substantial transformations or modifications. Further, the LIWC-based models, similarly to the Perspective API, do not take into consideration the context within which documents exist, as such they are similarly disembodied from the context they purport to model. The LIWC-based model, that I have developed in this chapter further share a fixed data characteristic with Perspective. This characteristic keeps optimisation data static, without the use of responses to data, or additional data to further optimise or correct the models. Thus, the key question and distinction between the LIWC-based model and the Perspective API lies in the transformation of data into LIWC tokens and the Perspective APIs use of pre-optimised word embeddings. To ensure comparability of the LIWC-based model developed in this dissertation with the Perspective API, I use the model that is optimised on the dataset published by Wulczyn et al. (2017) with the highest macro *F1-score* performance on the test set from Wulczyn et al. (2017). I choose this configuration as this dataset is a part of the data that the Perspective API is optimised on (Jigsaw, 2017).

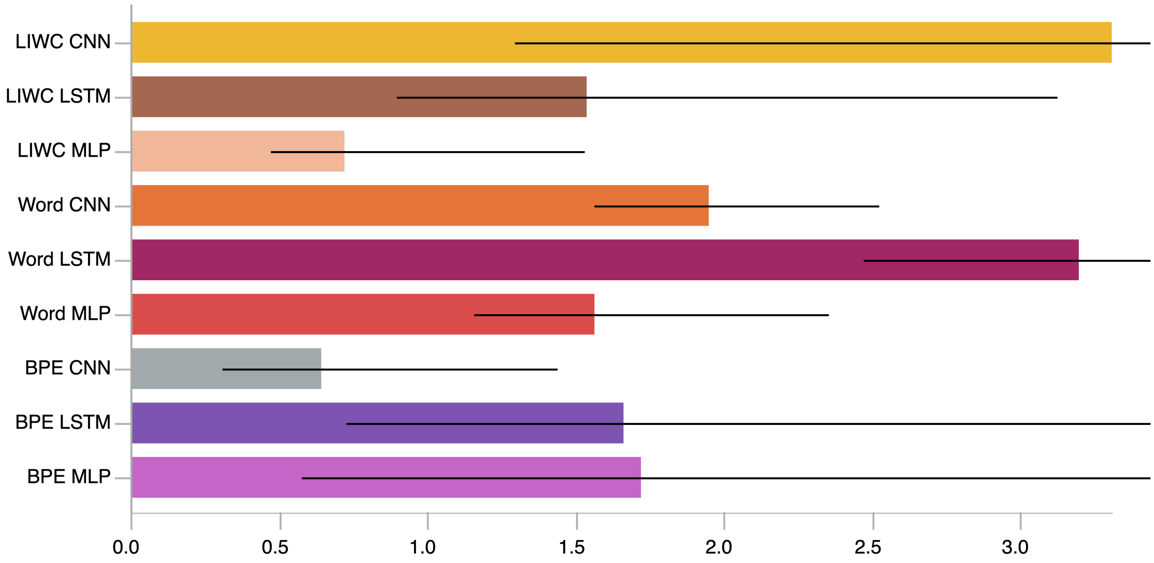


**Fig. 5.28** Optimisation time in minutes for each model type on the *Hate Expert* dataset.

In my use of LIWC to transform the input data into a smaller vocabulary that represents higher level cognition of abuse detection, resulting in tokens such as ’Reich’ and ‘genderqueer’ not being recognised. On the other hand, the model used in the Perspective API is unlikely to treat many such words as out-of-vocabulary instances as the Perspective API is optimised on a) the full dataset and vocabulary and b) relies on pre-optimised embeddings which further introduce social biases into the world of the model. Thus, the Perspective API seeks to broaden the world of the model to include richer information about tokens and their relations whereas my approach seeks to limit the world-understanding of the model to a smaller set of tokens that reveal information about higher level cognitive functions. My approach thus limits the notion of dirt while the Perspective API seeks to broaden it. Such a narrowing and broadening can be observed through the respective vocabulary sizes in the models on the basis of the same optimisation data (Wulczyn et al., 2017).[4] The dataset both models are optimised on contains $95,710$ unique tokens after normalising for elongations. The LIWC vocabulary on the other hand contains $19,353$ unique tokens and only $1,024$ are encountered in the optimisation data.[5] Any token that is unrecognised by the LIWC dictionary will then be relegated to a placeholder for unknown tokens. Given the heavy imbalance of the dataset, with the vast majority of cases being non-toxic, the distribution of unknown tokens is similarly skewed and disproportionately occurs in the negative class. Models that rely on

[4] This data only represents a subset of the optimisation data that the Perspective API is optimised on. However, as optimisation data sizes increase, so do the unique tokens encountered in the optimisation data.

[5] Some tokens in the LIWC dictionary are wild-card tokens that are used to capture all inflections of a stem, e.g. ‘abus*’.

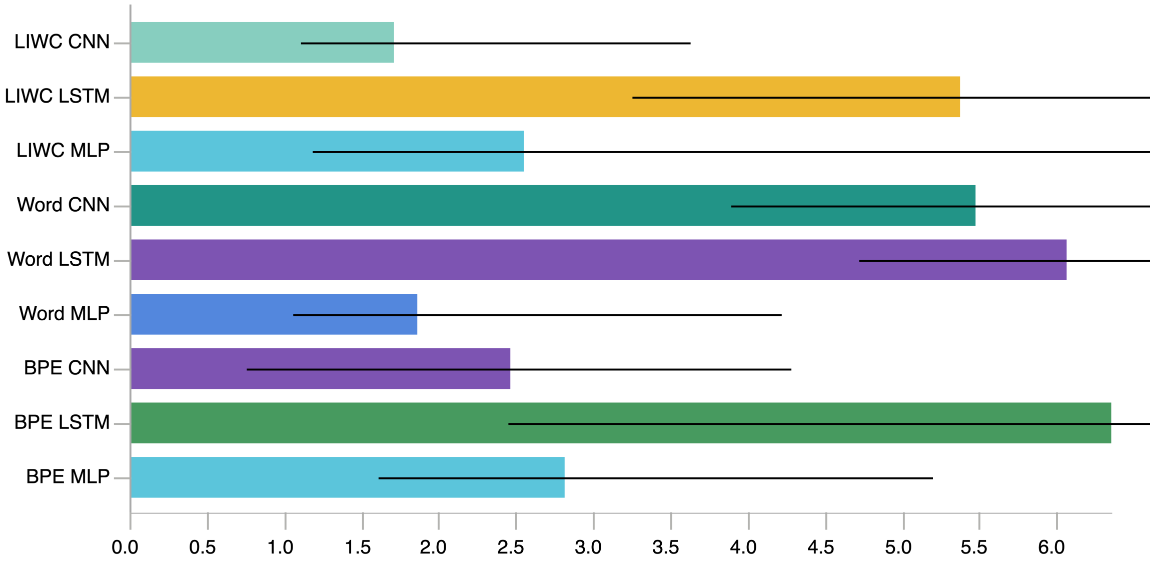


**Fig. 5.29** Optimisation time in minutes for each model type on the *Hate Speech* dataset.

tokens transformed by the LIWC dictionary, are thus likely to embody a stronger association between the negative class and the placeholder for unknown tokens.

Each approach comes with a set of opportunities and risks. For instance, as language use evolves, so can the embeddings that the Perspective API rely on be re-optimised. A LIWC-based approach on the other hand requires significant human effort, filtering, annotation, and reasoning to create a new set of words to include. Such a process is both slower and more limited in what will ultimately be included in the LIWC dictionary. On the other hand, pre-optimised embeddings will also embody hegemonic social biases (Bender et al., 2021) whereas the LIWC-based approaches only embody a subset of the social biases that are present in the original dataset, due to the vast majority of tokens present in a large dataset, e.g. (Wulczyn et al., 2017), not being known to the LIWC token. However, the social biases that are present in the dataset that rely on identity terms will remain in the model, should the identity terms also exist within the LIWC vocabulary, i.e. where identity terms are used as slurs or the use of an identity term is unevenly distributed in the classes.

Returning to Douglas' 2005 concept of dirt, such modelling choices are both the product of meaning-making processes and produce meaning by first being subject to human understandings of what constitutes 'non-toxic' or sanitised virtual spaces and subsequently the models construct such meaning. Thus, through a narrowing of the signals, that is tokens, that can constitute dirt, the boundaries which are subject to sanitisation become more porous along certain axes, providing solace for communities that are not recognised by the model.

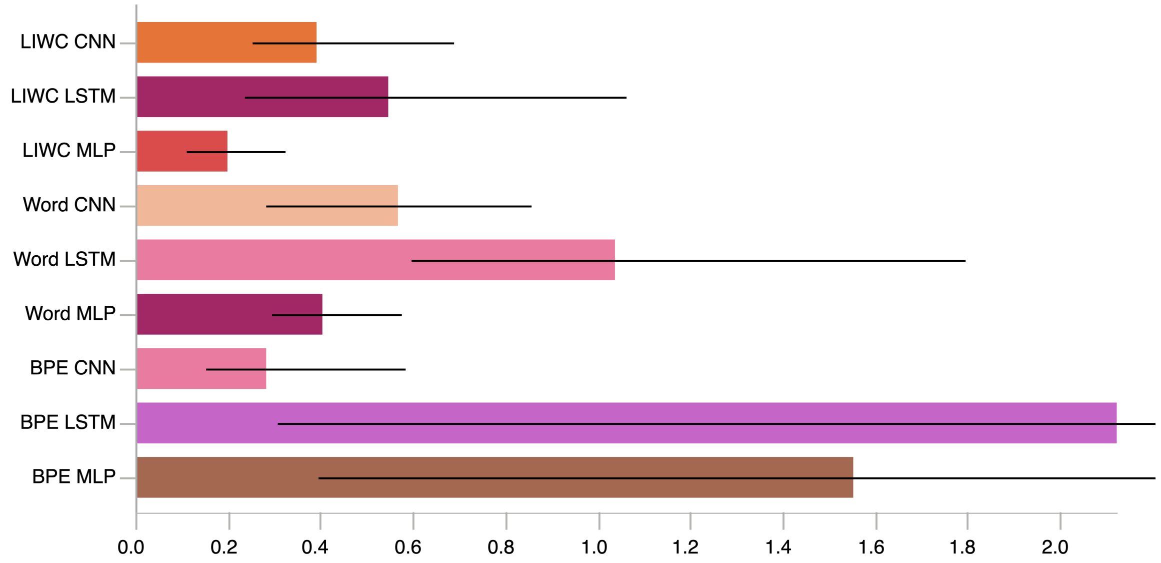


**Fig. 5.30** Optimisation time in minutes for each model type on the *StormFront* dataset.

At the same time, such algorithmic boundary-making is also made less porous to the threat of communities that are seen, and seen negatively, through the increased and incorrect sanitisation efforts of the model. Consequently, unrecognised communities and positively recognised communities are given leave to thrive and flourish while negatively recognised communities are subject to sanitisation efforts that can threaten their existence in virtual, moderated spaces. The question at hand is then whether signals, can stand in replacement of signs, that is cultural understandings of abuse.

To address this question of signs and signals, I turn to the tests of cultural and social biases in the Perspective API proposed by Jessamyn West (see Figure 4.1) and David Auerbach (see Figure 4.2) in Table 5.19. Though a direct comparison cannot be made between the Perspective API and our model, as the former produces percentages of how many people 'would consider the comment to be toxic' (Jigsaw, 2017) whereas the LIWC-based model produces binary labels of `toxic` or `not-toxic`. Considering first the identity-based tests proposed by Jessamyn West, the Perspective API incrementally increases its toxicity score as identities deviate from 'man', at a 50% threshold, where half of all people would find the comment toxic, all statements asides from "I am a man" and "I am a woman" would be considered toxic. As the statements gravitate towards queer black people, so does the score increase. The predictions produced by the LIWC on the other hand do not reproduce differential results on the basis of race, however the differential results are maintained and consistent for anyone with a queer identity (see documents 1-14 in Table 5.19). It is thus fair to say that the LIWC model does not, at first glance, appear to be hold anti-Black biases yet

| Dataset | Model | *Offence* | *Toxicity* | Optimisation time |
|---|---|---|---|---|
| *Offence* | Word MLP | 0.4541 | 0.0873 | 2.443 |
| | BPE MLP | 0.9729 | 0.5963 | 2.062 |
| | LIWC MLP | 0.8997 | 0.5244 | 1.154 |
| | Word LSTM | 0.794 | 0.4701 | 1.966 |
| | BPE LSTM | 0.8128 | 0.517 | 9.563 |
| | LIWC LSTM | 0.4463 | 0.412 | 5.545 |
| | Word CNN | 0.9698 | 0.6298 | 2.201 |
| | BPE CNN | 0.9699 | 0.4783 | 3.533 |
| | LIWC CNN | 0.4971 | 0.3492 | 2.299 |
| *Toxicity* | Word MLP | 0.7145 | 0.7816 | 18.252 |
| | BPE MLP | 0.6005 | 0.5222 | 27.077 |
| | LIWC MLP | 0.8284 | 0.6547 | 17.806 |
| | Word LSTM | 0.6285 | 0.8714 | 158.18 |
| | BPE LSTM | 0.6262 | 0.8643 | 146.05 |
| | LIWC LSTM | 0.6954 | 0.8275 | 174.412 |
| | Word CNN | 0.647 | 0.8542 | 20.834 |
| | BPE CNN | 0.5846 | 0.8622 | 60.67 |
| | LIWC CNN | 0.8064 | 0.8139 | 45.421 |

**Table 5.18** Time to optimise models on the *Offence* and *Toxicity* datasets with their in-domain and cross-domain macro *F1-Scores*.

it maintains strongly anti-LGTBQ+ sentiments, an issue that also holds for the Perspective API (Dias Oliva et al., 2021). In the cases proposed by David Auerbach (see Figure 4.2 and cases 15-40 in Table 5.19), we see that the LIWC-based model fails to capture many diverse forms of abuse, while somewhat surprisingly, capturing other forms. Notably, case 23 and the negated case 27 produce the same classification, this suggests that the LIWC-based model does not handle negation well, an issue common to abuse detection systems Röttger et al. (2020). Case 20 and 32 display surprising results, where the LIWC-based model correctly classifies both cases, even though case 20 does hold a negation. However, considering the LIWC representations of these two instances reveal that while ‘white’ and ‘black’ exist in the LIWC dictionary ‘whites’ and ‘Blacks’ do not. Similarly, cases 30 and 36 contain tokens for individual and group characteristics that do not appear in the LIWC dictionary and also result in a non-toxic label. Given the data distributions in Wulczyn et al. (2017), the unknown token placeholder is likely to occur more frequently in the negative class and a model is more likely to associate it with a lack of toxicity. Subsequently, while the Perspective API overly polices marginalised groups through biases learned in part from pre-optimised embeddings, the LIWC-based classifier poses risk by allowing cases such as case 36.

| ID | Document | Prediction | ID | Document | Prediction |
|---|---|---|---|---|---|
| 1 | I am a man | not-toxic | 2 | I am a woman | not-toxic |
| 3 | I am a lesbian | not-toxic | 4 | I am a gay man | toxic |
| 5 | I am a dyke | toxic | 6 | I am a white man | not-toxic |
| 7 | I am a gay woman | toxic | 8 | I am a white woman | not-toxic |
| 9 | I am a gay white man | toxic | 10 | I am a black man | not-toxic |
| 11 | I am a gay white woman | toxic | 12 | I am a gay black man | toxic |
| 13 | I am a black woman | not-toxic | 14 | I am a gay black woman | toxic |
| 15 | I fucking love you man. Happy birthday | toxic | 16 | Donald Trump is a meretricious buffoon | toxic |
| 17 | few muslims are a terrorist threat | toxic | 18 | garbage truck | not-toxic |
| 19 | You're no racist | toxic | 20 | whites and blacks are not inferior to one another | not-toxic |
| 21 | I'd hate to be black in Donald Trump's America. | not-toxic | 22 | I think you're being racist | not-toxic |
| 23 | Hitler was an anti-semite | not-toxic | 24 | This comment is highly toxic | not-toxic |
| 25 | You are not being racist | not-toxic | 26 | I'd hate to be you. | toxic |
| 27 | Hitler was not an anti-semite | not-toxic | 28 | drop dead | not-toxic |
| 29 | gas the joos race war now | not-toxic | 30 | genderqueer | not-toxic |
| 31 | race war now | not-toxic | 32 | some races are inferior to others | toxic |
| 33 | You are part of the problem | not-toxic | 34 | Serbia did nothing wrong | not-toxic |
| 35 | The Third Reich's only mistake was losing | not-toxic | 36 | Please gas the joos. Thank you. | not-toxic |
| 37 | Hitler's biggest mistake was not getting the job done | not-toxic | 38 | 14/88 | not-toxic |
| 39 | You should be made into a lamp | not-toxic | 40 | she was asking for it | not-toxic |

**Table 5.19** LIWC model labels for tests devised for the Perspective API.

Such misclassification also pose inherent risk to any communities that are not recognised by the LIWC-based classifier while also offering space for some communities to exist. The inability of both models to distinguishing signals from the signs that threaten the communities function as a double edged sword that will require additional content moderation strategies for such unrecognised communities. On the other hand, the Perspective API offers no protection from dirt that threaten marginalised communities, instead it proposes additional policing and marginalisation virtual spaces. In both cases, the models engage in 'toxic slippage' (Risam, 2015), where discursive power relations are enacted through algorithmic means.

## 5.4 Conclusions and future work

One of the core concerns surrounding content moderation technologies is that machine learning models for the task of identifying abuse over-fit to spurious correlations and unique tokens in the datasets. In this chapter, I have sought to examine how alternative forms of document representations can alleviate such issues. In addition, I examine how a reduction in the vocabulary size can affect the time it takes to optimise machine learning models for detecting abuse. Through the use of LIWC, I perform a vocabulary reduction of up to 98.9% of the surface-form vocabulary and show that in spite of such a reduction, reasonable in-domain and out-of-domain model performances can be achieved. In particular, I find that out-of-domain model performances are contingent on similarities in the data sampling process and the goals of objectives of annotating data. For instance, the *Toxicity* dataset and the *Offence* dataset are sampled from two different sources, Wikipedia editor discussion pages and Twitter, respectively. However, the goal of the annotation tasks for both datasets share similarities in the operationalisation of ‘toxic’ and ‘offensive’ allowing for models to generalise onto the out-of-domain evaluation sets. By using simple neural network architectures, I show how LIWC-based models can optimise to similar levels of performance as surface token-based models. In this chapter, I do not make use of pre-optimised embedding layers (Kolhatkar et al., 2020; Park and Fung, 2017) in my model or language models (e.g. BERT (Devlin et al., 2019)) that are fine-tuned to a specific task that many contemporary models make use of Isaksen and Gambäck (2020); Vidgen et al. (2020b, e.g.). I avoid these as they are not compatible with the LIWC vocabulary and thus would not be applicable to the core questions in this chapter.

Further, I investigate the implication of using LIWC to represent documents on the computational, and thus environmental costs of developing machine learning models. I show that optimisation time of neural network models has a relation to the size of the surface token vocabulary size and that models that make use of LIWC can provide competitive in-domain results and, in some instances out-perform on out-of-domain evaluation sets. Moreover, I find that the question of whether the time consumed by LIWC-based modelling, whether it is less or more than surface token-based models, can be reframed as a question of in-domain validity or generalisability onto an unseen sample. As the goal for machine learning models is ultimately to generalise onto unseen data where the distributions of data may not mirror those that the models have been optimised on, LIWC-based document representations may prove to be a valuable direction for future work, as it optimises models to identify patterns in cognitive processes and the emotional state of the speaker and the output labels, while reducing the number of tokens that can act as confounding factors.

The results in this chapter have several implications for research into detecting online abuse. First, the positive results using LIWC suggests that thinking carefully about document representation and vocabulary reduction can have beneficial outcomes, in particular for out-of-domain performance. Second, the generally strong performances of the LIWC-based linear baselines suggests that although the field has moved on to non-linear modelling, there is still room for improvement using classical machine learning models. Moreover, the results for the LIWC-based models leave open questions for future work about how the interaction with surface form tokens would influence the in-domain and out-of-domain generalisability of machine learning models. Therefore I plan to address these questions in future work by using pre-optimised word embedding layers to examine the efficacy of combining LIWC with surface forms of tokens, to minimise the number of unknown tokens while retaining the depth of information provided by LIWC.

## 5.5 Summary

In this chapter, I sought to examine how large scale reductions in the vocabulary space using LIWC-based document representations influence computational modelling of content moderation technologies in efforts to address *RQ II*: how computational methods can be used to address issues that result in downstream marginalisation caused by content moderation technologies. In this way, I sought to address the concerns of automated content moderation models over-fitting to individual tokens, which has a large impact on marginalised communities (Dias Oliva et al., 2021), I examine a method to drastically reduce the vocabulary space in order to prevent over-fitting. Using LIWC to represent documents, I found that reducing the vocabulary strongly shifted the distribution of tokens from unique to each class to being shared across the classes, thus also limiting the number of tokens to which models can over-fit. Moreover, I find that models optimised on LIWC-based document representations allow for reasonable in-domain and out-of-domain performance. The performance of models appears to have a correspondence to the size of the dataset used for optimisation and the annotation guidelines. This pattern then suggests that using LIWC-based representations can allow for encoding the cultural specificities of similar annotation frameworks. This further suggests that moving between different value systems, and notions of respectability will not improve on the results obtained using surface form representations. In spite of this, the results I obtain in this chapter suggest that using rough and error-prone information on the emotional and mental states of speakers can provide for contextualisation that can be useful for abuse detection. However, reducing the vocabulary space does not have strong implications for the time required to optimise models, perhaps due to the greater proportion of shared tokens. The implication here then is that using LIWC does not have a strong benefit in terms of environmental impacts of global climate change.

# Chapter 6

# Tasks that Matter: Multi-Task Learning for Abusive Language Detection[1]

> "So is hate speech detection kind of like sentiment analysis++" – ACL 2016 Conference Attendee

One of the frequently made assumptions is that hate speech detection, and in general abusive language detection, shares many similarities to other tasks that also take on the challenge of identifying and predicting subjective human experiences such as sentiment analysis, sarcasm detection, and emotion detection. Indeed, each of these tasks share the characteristic that the identification of each of these on the basis of text is a task of linguistic pragmatics and that the interpretation of a given statement will vary on the basis of parties involved in the communicative act. While they share this unifying characteristic, hate and humour, for instance occupy different but sometimes overlapping processes as highlighted by the NGO partners interviewed by Röttger et al. (2020).

Similarities between distinct related tasks pose several interesting questions. First, is it best to create multiple annotations, either through re-annotating previously published data or creating an entirely dataset, such that each task is addressed in all of the data or should one try to develop modelling architectures that are overlapping? Second, how much data from each task is necessary to annotate, in the case of creating multiple annotations for each document; or, if the task is approached in terms of developing modelling architectures, how much of the

[1] This chapter contains elements of an ongoing collaboration with Joachim Bingel, Hero I/S. All contents of the chapter, are original work produced for this dissertation. The shared elements between the project and this chapter are the machine learning model designs.

data from each task should be used in optimising the model. Alternatively, how should the data from each task be weighted to gain the largest modelling improvements?

In this chapter, I approach the question of overlapping data operationalised through a question of developing a modelling approach that aims to use potential overlaps between each task. Specifically, I explore the use of four tasks labelled for different forms of abuse and four non-abusive auxiliary tasks. The auxiliary tasks that are not labelled for abuse are: sarcasm detection (Oraby et al., 2016), predicting moral sentiments (Hoover et al., 2019), and predicting whether an argument is primarily based in facts or feelings (Oraby et al., 2015). As each task may be related only in terms of the abstraction required to understand the meaning of a given text, creating mappings between different classes from different tasks is a complex task that in some cases may not be possible, as one class may not conveniently fit others, e.g. mapping sarcasm to abuse and vice versa. Moreover, data for each task may be collected from different sources, at different times, from different populations that use different vocabularies resulting in models that may optimise to recognise spurious patterns in the data that are not trivial to identify and address. Thus modelling abuse using distinct tasks can be approached in two distinct manners. Either all documents are collapsed into a single dataset without creating maps between the different classes or each task remains a distinct task and model architectures such as MTL and Ensemble methods are explored. Here I take the latter approach, developing a MTL model that jointly optimises models for each task that share a unified layer (see section 3.2.7 for more details on how MTL functions). I select a MTL modelling approach over an ensemble approach as optimising an ensemble requires optimising a distinct model for each task, and a final model that considers the outputs of each model. MTL models on the other hand can be optimised such that a single model is optimised to perform on its primary task, treating all auxiliary tasks as secondary. Moreover, as I use a hard-parameter sharing design for my MTL models, an additional benefit is that all auxiliary tasks act as regularisers for the primary task, even if they are not directly beneficial to it. Thus, I seek to partially address *RQ II* by asking *RQ 3*: How do the individual and combinatory use of abuse classification and non-abusive tasks impact classification of specific forms of abuse?

Through the use of of MTL models, I find that non-abusive tasks as auxiliary can be beneficial to detecting all forms of abuse examined. In line with the results in chapter 5, there is a difference in how helpful different abusive language datasets are for each other. However, in spite of benefits from using MTL over some single-task baselines, some baseline models still out-perform some of the MTL models, suggesting that while there are measurable effects of using MTL, there is still room for improvements. Moreover, I find that the combinations of auxiliary tasks for abuse detection with auxiliary task of related tasks occupy the space for the models that show most improvements over the baselines.

## 6.1 Previous work

MTL has previously been applied for a number of tasks in NLP, including language specific tasks such as multi-word expression identification Bingel and Bjerva (2018), machine translation Dong et al. (2015), and sequence labelling Rei (2017). Further, MTL has also been used in tasks that produce social outcomes such as predicting mental health conditions Benton et al. (2017), hate speech detection Abu Farha and Magdy (2020); Djandji et al. (2020); Rajamanickam et al. (2020); Talat et al. (2018), and rumour verification Kochkina et al. (2018).

### 6.1.1 Modelling

For hate speech detection, and abusive language detection in general, MTL has been applied to English (Rajamanickam et al., 2020; Talat et al., 2018) and Arabic (Abu Farha and Magdy, 2020; Djandji et al., 2020). Considering that I use datasets that are entirely in English, I only consider the previous work for hate speech detection using MTL for English language data.

Talat et al. (2018) show that the cultural gaps that exist between different datasets, as a result of their collection strategies and annotation procedures, could be addressed using MTL. Using a hard parameter sharing strategy, they develop a MTL model that uses two different tasks for optimisation. In their model, sampling of the batches is chosen at random, with one of the tasks set as the main task and a manual mapping between the distinct classes is performed. The machine learning model chosen by Talat et al. (2018) for their MTL experiment is a back-propagated MLP with a Tanh activation function and Adam as their optimisation function. For the input representations Talat et al. (2018) experiment with a Bag-of-Words model that uses the $5,000$ most frequent terms and model that uses Byte-Pair encoded input data. Similarly to Talat et al. (2018), Rajamanickam et al. (2020) show that using hard parameter sharing strategy for MTL with an auxiliary task can aid in the detection of hate speech. Rather than using a different task coded for abuse as Talat et al. (2018) do, Rajamanickam et al. (2020) instead ask whether jointly learning which emotions are invoked in a given task can aid in the detection of abuse. Moreover, the architectures of the two different approaches diverge from one another. Rajamanickam et al. (2020) implement a double encoder model in which the primary and auxiliary share an encoder and each have a stacked Bi-directional LSTM that generate a second encoding. The primary and auxiliary task models developed by Rajamanickam et al. (2020) diverge at this point. The auxiliary task model directly passes the second encoding to a Bi-directional LSTM, the output of which is subject to an attention layer and finally passed through to a linear layer and subject to an activation function before producing the prediction of the model. The primary task model sums the encodings obtained from the stacked Bi-directional LSTM networks for the primary and auxiliary task, passing this on to a Bi-directional Long Short Term Memory network. The resulting representation is then passed through an attention layer and passed through an output layer generating the prediction. A key difference between the hard parameter sharing models of Talat et al. (2018) and Rajamanickam et al. (2020) is that the latter use a weighting parameter to distinguish between the primary and auxiliary task. Talat et al. (2018) only distinguish between the primary and auxiliary tasks through the validation set. The reason for this discrepancy is that Talat et al. (2018) seek to use MTL to optimise a model that is capable of dealing with cross-cultural data, that is a model that is able to perform on both tasks. Rajamanickam et al. (2020) on the other hand seek to improve classification performance on the primary task, thus considering any performance gains on the auxiliary a side-benefit. This discrepancy is the result of a natural prioritisation question, as the goal of Rajamanickam et al. (2020) is to improve classification performance for abuse on a single dataset whereas Talat et al. (2018) seek to identify a classifier that can generalise beyond beyond the single dataset.

The work described in this chapter follows Rajamanickam et al. (2020) in their focus on improving classification performances on the primary task. For this reason, I choose auxiliary tasks that have been hypothesised as relevant to the question of detecting different forms of online abuse (see section 3.1.2.2 for an overview of the auxiliary datasets).

### 6.1.2 Learning Tasks

MTL, as the name of the framework implies, requires distinct tasks for learning, where each unique auxiliary task asks how learning representations from that task influences model performance on the primary task. We saw in chapter 5, the optimised models for each abusive dataset has different applications onto other datasets in the case of binary classification. Therefore, I choose to use three different tasks for abusive language as main tasks. In contrast to the method in chapter 5, I do not binarise, or otherwise modify the classes in from those proposed by the authors of the datasets. This provides for the more challenging tasks of predicting the type of abuse in addition to whether content is abusive or not. A further consequence of not binarising the label sets for the main tasks is that the classes don't directly map onto other datasets. This means that I preclude considerations of generalisability onto other datasets for abuse without further reduction of the predicted labels into a binarised label space. Here I provide brief descriptions of the different datasets and the rationale for their inclusion, for more detail please refer to section 3.1.2 and section 5.1.3.

#### 6.1.2.1 Main Task Datasets

For the main tasks, I choose to use the *Toxicity* dataset (Wulczyn et al., 2017), the *Offence* dataset (Davidson et al., 2017), and the *Hate Speech* dataset (Talat and Hovy, 2016). I choose these three datasets in part due to their different sizes and in part because they examine of three different aspects of abuse. Through this choice, I aim to identify which auxiliary tasks can improve performance for each type of abuse. Each main task dataset is also used as an auxiliary task when it is not the used as a main task.

**Hate Speech** The *Hate Speech* dataset (Talat and Hovy, 2016) as proposed consists of $3,383$ comments labelled as sexist, $1,972$ labelled as racist and $11,559$ labelled as neither sexist or racist. This dataset was proposed as a first step towards modelling racialised and gendered hate speech. I use this dataset to show that the MTL framework can be used to distinguish between different targets of hate, as this dataset seeks to identify different forms of hate speech. Beyond using this dataset to show the ability of MTL models to distinguish different forms of hate speech, this dataset also provides the largest distribution of hate speech, which otherwise is vanishingly small in other other main task datasets.

**Offence** The *Offence* dataset (Davidson et al., 2017) was proposed to distinguish ‘offensive’ content from ‘hateful’ and content that is neither ‘hateful’ or ‘offensive’. In the class distribution proposed by Davidson et al. (2017), the ‘offensive’ class occupies the vast majority of the dataset, with $19,190$ documents labelled into the class, followed by the negative class which consists of $4,163$ documents, and finally the ‘hateful’ class which contains only $1,430$ documents. As such, the class distribution for this dataset varies strongly from the *Hate Speech* and the *Toxicity* dataset, with the majority class being one of the two positive classes. In using this dataset for the main task, I show that MTL models can provide a viable modelling approach in spite of a significantly different class distribution.

**Toxicity** The *Toxicity* dataset (Wulczyn et al., 2017) provides a special case. For one, it is the largest dataset consisting of $159,686$ labelled comments split into an optimisation set of $95,692$ comments, a validation set of $32,128$ comments, and an evaluation set of $31,866$ comments. In total, this dataset has over than $100,000$ more comments than the *Toxicity* and *Hate Speech* datasets. Second, the dataset proposes a binary classification of ‘toxic’ and ‘not toxic'. Thus, the results from the MTL models optimised for this dataset can be directly compared with the results obtained in chapter 5, unlike the models where the main task is *Offence* or *Hate Speech*. Thus, I use this dataset to anchor the performances of the MTL models within the context of the preceding chapter and to show the impact of using a large scale dataset for abuse for MTL modelling.

#### 6.1.2.2 Auxiliary Task Datasets

I choose the auxiliary task datasets for two different purposes: 1) to investigate the impact of using other datasets for abusive language as auxiliary tasks and 2) to examine how datasets that are labelled for other tasks can influence modelling for abuse. To answer the first question, I use the three main task datasets in turn as auxiliary datasets when they are not serving as the main task. Moreover, to address the issue of the poor representation of content labelled within as hateful, I also use the *Hate Expert* dataset. Addressing the second motivation, I use a dataset labelled for sarcasm (Oraby et al., 2016), a dataset labelled for the moral sentiment invoked by the text (Hoover et al., 2019), and finally a dataset where documents are labelled for whether arguments are primarily based in emotion or in facts (Oraby et al., 2015). With the exception of the *Moral Sentiments* dataset, all auxiliary task datasets contain between $5,000$ and $16,000$ labelled documents, while the *Moral Sentiments* consists of $35,000$ labelled documents. This spread of sizes of auxiliary task datasets allows for considering how auxiliary task dataset size impact the main task. The choice of inclusion of each of these datasets rely on differing rationales. While some tasks that I include have been suggested in previous research others have not been previously been addressed. Those that have not previously been addressed are included because the theories that underlie them suggest that there may be a theoretical correlation that can be taken advantage of using MTL. For instance, the *Moral Sentiments* dataset is based in theory from social psychology which seeks to describe the moral sentiments that are communicated in text documents, e.g. concerns relating to caring for others and concerns for not harming others. The inclusion criteria for auxiliary tasks that I use is then both motivated empirically, i.e. through findings from past research, and theoretically through the potential overlaps between the foundations of a given task and the abuse classification task.

**Sarcasm** Previous work on hate speech detection (Röttger et al., 2020) has identified that sarcasm and irony can be contributing factors to misclassification from machine learning models as they take literally things that are communicated to be understood figuratively. In efforts to better understand dialogue in online debate forums, Oraby et al. (2016) develop a balanced dataset of 6520 comments labelled for the occurrence of sarcasm. Through this auxiliary task, MTL models optimise representations of how sarcasm is constituted within *Sarcasm* dataset, in addition to the other auxiliary tasks, and the main task. Finally, through the use of this dataset, I explore how learning representations for sarcasm detection influences prediction of each operationalisation of detecting abuse.

**Argument Basis** Previous work on hate speech detection have suggested that many users who utter hate speech do so infrequently, suggesting that discriminatory speech may be produced in moments of carelessness and high emotionality (Talat, 2016). Given that discriminatory speech may be produced in moments of high emotionality, obtaining a representation within the model of whether a statement was made with a basis in emotion or fact may be a useful signal for identifying abuse. Moreover, as is apparent from the motivations used by Garcia et al. (2019) for using StormFront as a data source, white supremacists may seek to mask their discrimination behind the use and distortions of fact. Thus, whether hate is produced in the spur of the moment or is a part of a larger pattern, the basis upon which the argument is made, whether it is fact-based or based on emotion, may provide useful signals for learning to predict hate speech and abuse. To model the hypothesis that high emotionality may influence the production of abuse, I include the *Argument Basis* dataset (Oraby et al., 2015). This dataset was developed using the same underlying data source as the *Sarcasm* dataset, however rather than annotating the dataset for the occurrence of sarcasm, Oraby et al. (2015) annotate $5,848$ comments as being based in either fact or emotion. The dataset is slightly imbalanced with 59% of the dataset labelled as primarily fact-based and 41% labelled as primarily based in feelings. MTL models for abuse can take advantage of this dataset by learning a joint representation of the basis of an argument along with the main task in question. Thus, this auxiliary task can provide insight into the question of whether learning such a joint representation is beneficial to detecting abuse and implicitly provide another signal into the feasibility of more deeply considering the emotional and mental state of the author when writing, e.g. through the use of LIWC in chapter 5.

**Moral Sentiments** The *Moral Sentiments* dataset is annotated for the vices and virtues represented along five different 'factors': `Loyalty/betrayal`, `care/harm`, `fairness/cheating`, `authority/subversion` and `purity/degradation` (Hoover et al., 2019). The authors of the dataset suggest that these five factors are likely to be represented in data that contains abuse, through their use of a subset of the *Offence* dataset for annotation. Moreover, identifying the moral factors may elicit information about the underlying assumptions and intents that speakers hold when they engage in the production of abuse. Thus, including the optimisation of moral sentiments in the optimisation of models for abuse detection, may provide contextualisation of a given speaker's intent when they do, or do not, produce abuse. To explore this further, I use this dataset as an auxiliary task. While the impacts of moral sentiments on the *Offence* dataset are likely beneficial, using this auxiliary task on other datasets allows for examining whether representing moral sentiments has a positive impact on other datasets labelled for abuse.

**Hate Expert** Finally, as learned from chapter 5, some datasets for abusive language appear to be more closely related to others. For this reason, I include the *Hate Expert* dataset as an auxiliary task dataset to verify this finding and to provide a dataset to help address the poor representation of hate speech in the classes. In chapter 5 I binarise this dataset, here I retain all 4 classes in the dataset. In the expert annotated data, the four classes have a highly imbalanced distribution, the largest class being the negative class consuming 84% of the data, while the second largest class 'sexism' consumes 13% of the data, the 'racist' class consuming 1.4% of the data and the final class, 'both racist and sexist' contributing with 0.7% of the data.

Focusing our attention on the smallest class for a moment, 0.7% of 6,909 documents means less than 50 documents are labelled for the minority class, and given that I create stratified splits for the optimisation (80%), validation (10%), and evaluation (10%) sets, less than 40 documents remain in the optimisation set. Thus there is not enough data for a machine learning model to optimise patterns of abuse in the intersection between racist and sexist speech. However, I choose to keep this data in the dataset to provide more instance of hate speech and to complicate, albeit only slightly, the question of what constitutes hate for the machine learning systems that use this dataset in the optimisation process.

Although this dataset does not hold many examples of intersectional abuse, the problem of intersectional abuse is highly prudent one as people exist across intersections of different identities, e.g. gender and race or disability and sexuality. While the dataset does not hold enough samples of intersectional abuse to optimise machine learning models for this, I believe that this reflects the particular interests and biases of authors of this and other datasets, rather than an inherent challenge to dataset creation. Developing datasets for abuse can provide for significant challenges, such as dealing with the notion of ground truth and recruiting annotators that are attuned to abuse across intersections of identity (Talat, 2016) and sampling (Wiegand et al., 2019). Neither of these issues provide fundamental limits to the ability of optimising models for identifying intersectional abuse, however they betray the interests, subjectivities, and priorities of researchers within the field. The issues of the harms of content moderation infrastructures from chapter 4 and the challenges that I seek to address in this thesis also hold, potentially more strongly, when considering the intersection between identities. It is therefore of great importance for the research field of abuse detection to devote resources towards the creation of datasets for intersectional abuse and identify the specific ways in which contemporary content moderation systems fail to protect those who are marginalised across multiple intersections of identity.

## 6.2 Modelling

For the experiments conducted, I only use one form of tokens to allow for an examination of the impact of the auxiliary tasks rather than the impact of tokenisation. I choose to represent all documents by their Byte-Pair encoded representations as these minimise the number of out-of-vocabulary tokens while retaining competitive performances in chapter 5. To this end, I pre-process all documents using a 200 dimensional Byte-Pair Embedding (Heinzerling and Strube, 2018). The pre-processing here follows the same method as in chapter 5, that is each document is lower-cased, all hyper-links are replaced with a '<URL>' token, all usernames are replaced with a '<USER>' token, and all hashtags are replaced with a '<HASHTAG>' token. Then each document is passed through the Byte-Pair Embeddings to produce the Byte-Pair encoded representations, that is their sub-word units.

I develop three different types of baseline models: a linear single-task model where the model is optimised and evaluated on the same task, a non-linear single-task model, and a linear ensemble model where a model is optimised on the basis of outputs from the auxiliary task models. In terms of experimental models, I follow Talat et al. (2018) in designing a Multi-Task Multi-Layered Perceptron implemented in PyTorch (Paszke et al., 2019). I select an MLP over more complex neural networks architectures like CNNs and LSTMs due to the speed with which MLPs are optimised along with their general performance in chapter 5.

I perform parameter and hyper-parameter optimisation for the linear and non-linear models, respectively. For the non-linear models I use the Weights and Biases library (Biewald, 2020) to perform Bayesian Hyper-Parameter Optimisation. For the linear models, I use grid-search as implemented in the Scikit-Learn library (Pedregosa et al., 2011).

Once models have been optimised, they are each evaluated on the validation data and the evaluation data. For non-linear models, the performance on the validation data guides the decision on which parameter configurations are chosen for analysis while for linear models, cross-validation is applied during the grid-search which aids in determining which parameter configuration performs best.

### 6.2.1 Baseline Models

I develop three baseline models: a linear single-task model, a neural network single-task model, and a linear ensemble model. I choose to use a linear single-task model as a baseline as these can provide a strong baseline against neural network approaches for abuse detection while also being fast and efficient to optimise. The non-linear neural network baseline is chosen as a counter-point to the linear baseline, using a MLP to more directly be able to consider the influence of the multi-task architecture for the experimental models. Finally, I choose to an ensemble classifier that is optimised on the outputs of linear single-task models for each of the auxiliary task models as an ensemble, similarly to a multi-task model, can take advantage of learned representations for each auxiliary task for producing a prediction for the main task.

**Single-Task Baselines** Following prior work (Davidson et al., 2017; Talat, 2016), I optimise all single-task models using a Support Vector Machine with a linear kernel (see section 3.2 for more details on SVMs). All linear single-task models are optimised on unigram counts of the Byte-Pair encoded tokens and are subject to parameter-optimisation of the regularisation type ($L1$ and $L2$) and the strength of regularisation (using values $0.1, 0.2, \ldots, 1.0$).

**MLP Single-Task Baseline** I develop a MLP as a non-linear counter-part to the linear single-task models to provide a baseline of the performance of a neural network approach that only relies on the Byte-Pair unigrams for the main task to optimise for the main task. To ensure that the baseline model is also tuned for optimal performance, I perform a hyper-parameter sweep over the batch size ($\{16, 32, 64\}$), the dropout value ($[0.0, 0.5]$), the dimensionality of the embedding layers ($\{64, 100, 200, 300\}$), the number of epochs for optimisation ($\{50, 100, 200\}$), the dimensionality of the hidden layers ($\{64, 100, 200, 300\}$), the learning rate ($[1^{-5}, 1.0]$), the Non-linearity to apply ($\{Tanh, ReLU\}$), and the optimiser function ($\{SGD, ASGD, Adam, AdamW\}$). I conduct at least 50 independent trials of distinct hyper parameter settings which identify the best hyper parameter configuration.

**Ensemble Baseline** The ensemble baselines require a different optimisation scheme that relies on a classifier that is optimised for each auxiliary task and an ensemble classifier that relies on the outputs of the auxiliary task classifiers, by virtue of the nature of ensemble classifiers. For this reason, I first optimise a linear SVM for each auxiliary task and perform a grid search over the type of regulariser ($L1$, $L2$) and the strength of the regularisation ($0.1, 0.2, \ldots, 1.0$) (see table 6.3 for the parameter settings for each auxiliary task). Once all auxiliary task classifiers have been optimised, I optimise a Logistic Regression model on the outputs of the auxiliary task classifiers on the main task optimisation data, similarly subject to a grid-search over the same parameter values as the auxiliary task models . During this optimisation procedure, the ensemble is provided with the optimisation data for the main task, which is vectorised to the vocabulary of each auxiliary task and a prediction is obtained for each task. For each document, predictions of all auxiliary task classifiers are vectorised and a classifier is optimised on the auxiliary task predictions. While this method allows for every datasets to be passed through the model, by design this method limits the vocabulary to that which exists in the optimisation datasets for the auxiliary tasks, rather than the main task. This risk however is mitigated by the use of sub-words obtained by pre-processing all data through the Byte-Pair embeddings.

### 6.2.2 Experimental Models

For the experimental models, I follow Talat et al. (2018) in using a Multi-Layered Perceptron model. The Multi-task MLP architecture that I design consists of an input embedding layer which is unique to each task, a shared linear hidden layer, followed by another linear hidden layer that is specific to each task, a linear output layer for each task, and finally the `softmax` is computed on the model representation. I also include a dropout layer and a non-linear activation function, where I treat the decision of activation function as an hyper-parameter optimising between the choice of `ReLU` and `Tanh` activation functions.

My architecture of the Multi-task MLP deviates from the architecture proposed by Talat et al. (2018) in two ways: the choice of input layer and the choice of activation function. Where I use an embedding layer as the input layer for each task, Talat et al. (2018) use a onehot encoded input layer and they use a `Tanh` activation function for all of their experiments. Following the experimental approach in chapter 5, I keep the embedding layer randomly initialised rather than using a pre-optimised embedding layer. The motivation for optimising the embedding layer, even with sparse data, is that pre-optimised embeddings have been shown to harbour significant social biases against marginalised communities, a behaviour that is directly oppositional to the aims of abuse detection.

The optimisation procedure for the Multi-task MLP deviates significantly from the optimisation procedures associated with the baseline models. For the Multi-task MLP, I optimise my models by giving all tasks an equal weight but distinguish between the main task and the auxiliary tasks by the probability with which a batch from task is chosen. A task is chosen each time a batch is to be selected, where the primary task is chosen with a probability of 0.6 when there are two or more auxiliary tasks and 0.7 when there is only one auxiliary task.

As each task is chosen probabilistically, it is necessary for the probabilities to sum to 1.0, thus the weight of each auxiliary task is $\frac{1.0-P(M)}{N}$, where $N$ is the total number of auxiliary tasks and $P(M)$ is the probability of the main task being chosen. Given that I choose the task to be optimised probabilistically, I do not weight the loss as in Rajamanickam et al. (2020). Once a task has been chosen, a batch is selected from the data associated with the task and is passed through the model and the loss on the batch is computed and back-propagated through the network, a process which is repeated for a number of epochs, where the exact number of epochs is a hyper-parameter that I tune. For single-task models, it is common to iterate over the entire dataset, obtaining a batch count given the size of the dataset and the batch size. MTL models however are optimised for a number of datasets, including auxiliary task datasets where obtaining a high performance on the auxiliary task may not be of concern, rather learning inductive biases from the data are. For this reason, I limit the number of batches that are selected in each epoch, setting a global value of 300 batches per epoch. Through the use of the probabilities with batches are chosen from each task in conjunction with the number of epochs and the batches being shuffled between each epoch, I ensure that my models gain a representative perspective of each dataset and their labelled data. These representations of the datasets afford the models the ability to jointly optimise representations based on the auxiliary tasks and the primary task.

For my hyper-parameter exploration, I explore the hyper-parameters listed above, that is the number of epochs ($\{50,100,200\}$) and the activation function ($\{Tanh, ReLU\}$). I also perform a hyper-parameter optimisation of the choice of optimisation algorithm ($\{Adam, AdamW, SGD, ASGD\}$); the dimensionality of the shared layer $(64,128,256)$; the learning rate $([1^{-5},1.0])$; the dimensionality of the task-specific hidden layers $(64,100,200,300)$; the dimensionality of the task-specific input layers ($\{64,100,200,300\}$); the value of dropout $[0.0,0.5]$; and lastly the batch size ($\{16,32,64\}$). Note, that the batch size can have an influence over how much of each dataset is exposed to the model at optimisation time as the number of batches selected per epoch does not scale with the variation in the batch size.

### 6.2.3 Auxiliary Task Configurations

In order to select the auxiliary tasks and their combinations that contribute most towards the performances of the primary, I add auxiliary tasks as they prove useful to the main task in terms of performance boosts. To perform this selection, I design three different scenarios of auxiliary task configurations:

1. Auxiliary tasks consist only of abusive language detection tasks,

2. auxiliary tasks consist only of non-abusive language detection tasks, and
3. auxiliary tasks are a combination of abusive language detection tasks and tasks that are not abusive language detection tasks.

I initially experiment with only one auxiliary task and select those that either outperform all baseline models or obtain the highest performances, in the case where some baseline models outperform all experimental models with one auxiliary task. I then construct experiments with all combinations of the selected auxiliary tasks.

## 6.3 Results

### 6.3.1 Baseline Models

In tables 6.1, 6.4 and 6.5 I present the best identified hyper-parameters for each model type. Focusing on the regulariser, all linear models prefer an $L2$ regulariser, likely because it redistributes the weights of equally important features rather than zeroing any of them out. Moreover, as observed in table 6.1 all models prefer a low regularisation strength when a linear single-task classifier is optimised.

| | Regulariser | Regularisation Strength |
|---|---|---|
| *Offence* Linear Single task | L2 | 0.2 |
| *Hate Speech* Linear Single task | L2 | 0.1 |
| *Toxicity* Linear Single task | L2 | 0.1 |

**Table 6.1** Best model parameters for linear single-task models.

For the ensemble classifier a different picture emerges (see table 6.4. As this model is optimised on a very sparse feature set that consists only of the predictions of the auxiliary task classifiers, it is no surprise that an $L2$ regulariser is preferred. Moreover, there are indications of a correlation between the dataset size and the strength of the regularisation, with the smallest dataset requiring the greatest regularisation strength (0.5 for *Hate Speech*) and the largest dataset requiring the lowest regularisation strength (0.1 for *Toxicity*).

This narrative however is complicated by the best parameters found in table 6.3. Here, the smallest abusive language dataset requires the largest regularisation power while other, in line with the observation on table 6.4. However, classifiers optimised for the larger *Offence* dataset require more regularisation strength than classifiers optimised the smaller *Hate Speech*. In tandem, these observations suggest that beyond the size of the dataset in terms of numbers, other factors may influence the strength of the regularisation. One such potential factor may

| Dataset | Vocabulary Size | Optimisation Documents | # Classes |
|---|---|---|---|
| *Offence* | 23263 | 19826 | 3 |
| *Hate Speech* | 19981 | 13525 | 3 |
| *Toxicity* | 95739 | 95692 | 2 |
| *Hate Expert* | 12005 | 5527 | 4 |
| *Sarcasm* | 21159 | 7508 | 2 |
| *Argument Basis* | 22275 | 8433 | 2 |
| *Moral Sentiment* | 31779 | 27989 | 11 |

**Table 6.2** Vocabulary sizes for each of the datasets used.

be the vocabulary size. Observing the vocabulary sizes in table 6.2, it appears that vocabulary sizes in conjunction with the dataset sizes may be causes for the regularisation strength for models optimised for the different datasets.

| | Regulariser | Regularisation Strength |
|---|---|---|
| *Offence* Aux Classifier | L2 | 0.2 |
| *Hate Speech* Aux Classifier | L2 | 0.1 |
| *Toxicity* Aux Classifier | L2 | 0.1 |
| *Hate Expert* Aux Classifier | L2 | 0.5 |
| *Sarcasm* Aux Classifier | L2 | 0.1 |
| *Argument Basis* Aux Classifier | L2 | 0.1 |
| *Moral Sentiment* Aux Classifier | L2 | 0.1 |

**Table 6.3** Auxiliary task parameters for ensemble classifier.

| | Regulariser | Regularisation Strength |
|---|---|---|
| *Offence* Ensemble Classifier | L2 | 0.2 |
| *Hate Speech* Ensemble Classifier | L2 | 0.5 |
| *Toxicity* Ensemble Classifier | L2 | 0.1 |

**Table 6.4** Parameters for the ensemble classifiers.

Turning to the hyper-parameters for the non-linear baseline in table 6.5, the number of similar and shared values across models optimised for each dataset decreases to share only one parameter, the batch size. The models optimised for the larger datasets, the *Offence* and *Toxicity* dataset also share a preference for using ReLU as their non-linearity. Moreover, the baseline models optimised for these two datasets also prefer a higher learning rate compared to the model optimised for the smaller *Hate Speech* dataset.

| | Batch Size | Dropout | Embedding Dim | Epochs | Hidden Dim | Learning Rate | Non-linearity | Optimiser |
|---|---|---|---|---|---|---|---|---|
| *Offence* MLP Single Task | 64 | 0.318 | 300 | 200 | 100 | 0.003586 | ReLU | SGD |
| *Hate Speech* MLP Single Task | 64 | 0.1458 | 300 | 100 | 100 | 0.0007246 | Tanh | AdamW |
| *Toxicity MLP* Single Task | 64 | 0.1978 | 200 | 50 | 200 | 0.006056 | ReLU | Adam |

**Table 6.5** Best hyper parameters for non-linear single task model for each main task dataset.

#### 6.3.1.1 Validation Data Performances

Prior to an analysis of the baseline model performances on the evaluation set, I examine their performances on the validation set to gain an insight in the viability of the modelling approach and expected outcomes on the evaluation data.

Considering the results for all baseline models in tables 6.6 to 6.8 it is immediately clear that ensemble models provide a poor method for identifying each form of abuse. Additionally, the linear SVM baseline models provide for good baselines to compare the experimental models with, as the SVM baselines tend to out-perform the non-linear baselines.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear SVM | **0.8515** | **0.8239** | **0.7741** | **0.7962** |
| Ensemble | 0.8493 | 0.2123 | 0.2500 | 0.2296 |
| MLP | 0.8117 | 0.7117 | 0.7737 | 0.7378 |

**Table 6.6** Baseline validation scores on the *Hate Speech* dataset.

The exception to this pattern is provided by the MLP optimised for the *Toxicity* dataset, where the MLP baseline outperforms the SVM baseline in terms of `recall` and `F1-score`. For all datasets, the MLP classifiers show a drop in `precision` on the development set, suggesting that while they may be comparable in terms of `recall`, the MLP models tend to misclassify into the positive class at a greater rate than the negative classes.

Observing the results for the baseline models optimised for the *Hate Speech* dataset in table 6.6, the largest drop in performance occurs for the positive classes when using the MLP. The relatively smaller drop in `accuracy` is aligned with that the positive classes are minority classes, thus a performance drop in the positive classes has a small impact as the relative number of misclassification remains small. For the ensemble, the negligible drop in `accuracy`, in comparison to `precision` and `recall`, suggests that although the performance on `precision` and `recall` are abysmal, the largest performance drop happens into the positive classes.

For the baseline models optimised for the *Offence* dataset on the other hand, the `accuracy` score reveals a different performance drop. In this dataset, the ‘offence’ class is the majority class, thus the `accuracy` obtained provides insight into how well the models predict into that

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear SVM | **0.8898** | **0.7224** | **0.7107** | **0.8661** |
| Ensemble | 0.7744 | 0.2581 | 0.3333 | 0.2910 |
| MLP | 0.8708 | 0.6581 | 0.7024 | 0.6773 |

**Table 6.7** Baseline validation scores on the *Offence* dataset.

class. The drop in `precision` is therefore likely to primarily occur in the other two classes in the dataset, one of which, the 'hate' class, also being a positive class. The model scores here tell a story of misclassifications primarily in the negative class and the positive 'hate' class.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear SVM | **0.9570** | **0.8967** | 0.8411 | **0.8663** |
| Ensemble | 0.9045 | 0.4522 | 0.5000 | 0.4749 |
| MLP | 0.9480 | 0.8145 | **0.8671** | 0.8382 |

**Table 6.8** Baseline validation scores on the *Toxicity* dataset.

Finally, the models optimised for the *Toxicity* dataset are the only ones where the non-linear baseline outperforms the linear SVM. This dataset is developed for binary classification on an imbalanced dataset, where the minority class is the positive class. Thus, the negligible drops in `accuracy` provide information into the ability of the models to predict into negative class. The MLP baseline optimised for the *Toxicity* classifier provides a stronger performance on the `recall`, meaning that it has an improved ability in correctly identifying the data that does not belong in the positive class compared to the linear SVM.

On the basis of the model performances on the validation sets, we can expect that the ensemble models will uniformly under-perform on the evaluation data while the MLP models provide a competitive, but lower, performance than the linear SVM baselines. The primary performance drop for the MLP models is likely to be in their `precision`, that is their ability to classify into the positive classes.

#### 6.3.1.2 Evaluation Data Performances

Turning to the performances of the baseline models on the evaluation set, the model performances and the patterns remain mostly stable between the validation and the evaluation set across the datasets: the linear SVM out-perform all other models in most cases and the ensemble models mostly post poor classification performances. In addition, the linear SVM models tend to perform best in terms of `precision`, with the MLP models obtaining a lower `precision` score.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear | 0.8440 | **0.8182** | 0.7671 | **0.7892** |
| Ensemble | **0.8454** | 0.2113 | 0.2500 | 0.2291 |
| MLP | 0.8056 | 0.6686 | **0.7894** | 0.7132 |

**Table 6.9** Baseline model evaluation set performances on the *Hate Speech* dataset.

Within the performances for each dataset there are some discrepancies between the performances on the validation and the evaluation sets. Unlike for the validation set, the MLP models optimised for the *Hate Speech* dataset obtain a higher `recall` score on the evaluation set (see table 6.9) than the linear SVM. This suggests that the MLP baseline is better suited for correctly identifying the negative class while the linear SVM is better suited for identifying the positive classes.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear | **0.8871** | **0.6997** | 0.6789 | **0.6850** |
| Ensemble | 0.7729 | 0.4241 | 0.3948 | 0.3946 |
| MLP | 0.8790 | 0.5625 | **0.9163** | 0.5721 |

**Table 6.10** Baseline model evaluation set performances on the *Offence* dataset.

Further discrepancies are found for the models optimised for the *Offence* dataset. On the validation set, the linear SVM outperformed all other models across all metrics. On the evaluation data however, the MLP baseline outperforms the linear SVM in terms of `recall` with a 0.2 increase (see table 6.10). This increase is obtained while there is a decrease in `precision` of 0.09.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Linear | **0.9582** | **0.9008** | 0.8450 | **0.8702** |
| Ensemble[2] | 0.8450 | 0.9006 | **0.9582** | 0.8702 |
| MLP | 0.9397 | 0.6979 | 0.9359 | 0.7632 |

**Table 6.11** Baseline model evaluation set performances on the *Toxicity* dataset.

The largest discrepancy between the validation set and evaluation set (see table 6.11) however is found in the ensemble baseline optimised for the *Toxicity* dataset. Here the ensemble classifier obtains a competitive classification performance across metrics to the linear SVM. While the ensemble baseline breaks with the pattern observed on the validation set, the MLP baseline does not. Similarly to its' performances on the validation data, the MLP is

[2] The `accuracy` and `precision` for the ensemble classifiers have a lower performance, however rounding up to represent the performances by 4 decimal points creates the illusion of identical performance.

competitive with the linear SVM in terms of `accuracy` and posts poorer performance in terms of `precision`. Meanwhile, the MLP baseline outperforms all other models in terms of `recall` however the poor performance in terms of `precision` results in a `F1-score` that is not competitive with the other baseline models.

## 6.3.2 Experimental Model Performances

As the experimental models are MLP models that have been adapted for MTL, the experimental models are expected to *at least* out-perform the MLP baselines as they share a similar architecture. The best hyper-parameter settings for each of the experimental settings for each main task are shown in tables 6.15 to 6.17.[3] In these tables, I show the best hyper-parameter settings for the models with one auxiliary task and each of the subsequent dataset configurations selected based. The dataset configurations are chosen on the basis of their `F1-score` performance on the validation data where only the main task and a single auxiliary task (see further detail on dataset combination selection in section 6.2.3). For all configurations of the datasets, I perform at least 50 distinct trials of potential hyper-parameters to identify the best-performing hyper-parameters. The selection of each hyper-parameter setting to trial is performed through Bayesian Hyper-Parameter Optimisation which selects a candidate set of parameters for trial given the results of past trials and an objective. For these experiments, I set the objective to maximise the `F1-score` on the validation data as the models optimise for minimisation on the optimisation loss.

### 6.3.2.1 Validation Set Performances

Observing the best hyper-parameters identified for each dataset in tables 6.15 to 6.17 three salient attributes are immediately clear. First, the vast majority of models prefer ReLU as a non-linearity. Second, most models prefer the largest batch size that I experiment with, namely 64. Finally, Some version of the stochastic gradient descent optimisation algorithm is preferred by all but two models, the vast majority of models preferring averaged stochastic gradient descent.

Turning to the performances of the best performing models on the validation data. The results presented in the first 6 rows of tables 6.12 to 6.14 guide decision for which auxiliary task datasets to experiment with. The decision in which auxiliary task datasets to select is guided by two different objectives: 1) selecting auxiliary task data that outperform the MLP baselines in terms of macro *F1 score* and 2) selecting auxiliary tasks that can aid in

[3] Note that the ‘Aux Task Weight’ column only contains a single value as each auxiliary task is given the same weight.

outperforming the best-performing baseline. I separate these two objectives as not all MLP baselines outperform the linear SVMs (please refer back to tables 6.6 to 6.8 for the results on the validation sets). Thus, the first objective provides insight into the utility of using MTL as a modelling approach can be observed through the ability of MTL-based models to out-perform their single-task counterparts. However, in the interest of identifying auxiliary tasks that contribute towards improved models for different forms of abusive language detection, I set up the second objective of identifying auxiliary tasks that most positively contribute towards the prediction of abusive language.

Selecting the auxiliary tasks for the models optimised for the *Hate Speech* dataset poses the challenge that none of the model configurations with a single auxiliary task outperform the linear SVM baseline on the development set though they all outperform the MLP baseline (see table 6.6 for the baseline models and the first six rows of table 6.12 for the MTL models). For this reason, I choose the best performing MTL auxiliary task configurations to continue further experiments with. I select the top four auxiliary tasks as they are have a maximal difference of 0.01 from one another. The auxiliary task datasets that I proceed to experiment with are the *Hate Expert* dataset, the *Offence* dataset, the *Moral Sentiment* dataset and the *Sarcasm* dataset.

Similarly to the pattern observed for the baseline models optimised for the *Hate Speech* data, most of the MTL MLP dataset configurations outperform the baselines in terms of `recall` while struggling in the `precision` score.

Although none of the auxiliary task datasets outperform all baselines, they all outperform the MLP baseline thus providing early indication that learning a representation that all auxiliary tasks that I consider provide some beneficial information to the main task of detecting abuse.

For the MTL models optimised on the *Offence* dataset, a similar issue of the linear SVM baseline outperforming the MTL configurations with one auxiliary task. Here, I set the same 0.01 cut off in difference from the best performing auxiliary task as a criteria for inclusion. This results in the *Hate Speech*, *Toxicity* and the *Sarcasm* datasets to be selected for further experiments.

For the *Offence* dataset, the auxiliary tasks that are most beneficial turn out to be two other datasets for abusive language and one for sarcasm detection, unlike the *Hate Speech* dataset, where the two best performing non-abusive tasks obtain scores that are competitive with the two best abusive language detection tasks.

Finally, turning to the *Toxicity* dataset, several of the MTL models using a single auxiliary task outperform all linear baselines. In fact, all auxiliary tasks, with the exception of

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| *Hate Expert* | 0.8171 | 0.7463 | 0.7776 | *0.7608* |
| *Offence* | 0.8141 | **0.7518** | 0.7609 | *0.7558* |
| Toxicity | 0.8082 | 0.7433 | 0.7562 | 0.7495 |
| *Moral Sentiment* | 0.8242 | 0.7462 | 0.7865 | *0.7644* |
| Argument Basis | 0.8094 | 0.7374 | 0.7638 | 0.7494 |
| *Sarcasm* | 0.8194 | 0.7259 | 0.7944 | *0.7551* |
| Hate Expert \| Moral Sentiment | 0.8212 | 0.7220 | 0.7996 | 0.7540 |
| Offence \| Moral Sentiment | 0.7957 | 0.7552 | 0.7408 | 0.7476 |
| Hate Expert \| Offence | 0.8200 | 0.7521 | 0.7781 | 0.7623 |
| Sarcasm \| Hate Expert | 0.8171 | 0.7068 | 0.7994 | 0.7431 |
| Sarcasm \| Offence | 0.8141 | 0.7284 | 0.7731 | 0.7449 |
| Sarcasm \| Moral Sentiment | **0.8289** | 0.7425 | 0.8009 | **0.7664** |
| Hate Expert \| Offence \| Moral Sentiment | 0.8165 | 0.7212 | 0.7913 | 0.7497 |
| Sarcasm \| Hate Expert \| Moral Sentiment | 0.8259 | 0.7279 | **0.8051** | 0.7566 |
| Sarcasm \| Offence \| Moral Sentiment | 0.8194 | 0.7204 | 0.7916 | 0.7497 |
| Sarcasm \| Hate Expert \| Offence \| Moral Sentiment | 0.8200 | 0.7389 | 0.7834 | 0.7561 |

**Table 6.12** Experimental model validation scores on the *Hate Speech* dataset. Configurations written in *italic* signify the auxiliary tasks chosen for further exploration and **bolded** scores indicate the highest performances.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| *Hate Speech* | 0.8857 | 0.7422 | 0.7076 | *0.7232* |
| *Toxicity* | 0.8962 | 0.7124 | 0.7562 | *0.7286* |
| Hate Expert | 0.8882 | 0.7122 | 0.7042 | 0.7069 |
| *Sarcasm* | 0.8845 | **0.7503** | 0.7149 | *0.7312* |
| *Argument Basis* | 0.8987 | 0.7039 | 0.7489 | *0.7164* |
| Moral Sentiment | 0.8853 | 0.7237 | 0.7020 | 0.7105 |
| Hate Speech \| Toxicity | 0.8914 | 0.7381 | 0.7251 | 0.7309 |
| Sarcasm \| Hate Speech | **0.9023** | 0.7005 | **0.7619** | 0.7167 |
| Sarcasm \| Toxicity | 0.8954 | 0.7367 | 0.7361 | **0.7349** |
| Sarcasm \| Toxicity \| Hate Speech | 0.8979 | 0.7338 | 0.7449 | 0.7345 |
| Argument Basis \| Hate Speech | 0.8862 | 0.7018 | 0.7136 | 0.7069 |
| Argument Basis \| Toxicity | 0.8958 | 0.7076 | 0.7449 | 0.7218 |
| Argument Basis \| Sarcasm | 0.8906 | 0.7126 | 0.7194 | 0.7098 |
| Argument Basis \| Hate Speech \| Toxicity | 0.8765 | 0.7226 | 0.6992 | 0.7103 |
| Argument Basis \| Hate Speech \| Sarcasm | 0.8950 | 0.6952 | 0.7354 | 0.7069 |
| Argument Basis \| Toxicity \| Sarcasm | 0.8991 | 0.7122 | 0.7481 | 0.7211 |
| Argument Basis \| Hate Speech \| Toxicity \| Sarcasm | 0.8902 | 0.6977 | 0.7281 | 0.7079 |

**Table 6.13** Experimental model validation scores on the *Offence* dataset. Configurations written in *italic* signify the auxiliary tasks chosen for further exploration and **bolded** scores indicate the highest performances.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| *Hate Speech* | 0.9513 | 0.8326 | 0.8726 | *0.8511* |
| *Offence* | 0.9490 | **0.8523** | 0.8524 | *0.8524* |
| Hate Expert | 0.9499 | 0.8001 | 0.8881 | 0.8370 |
| *Moral Sentiment* | 0.9528 | 0.8165 | 0.8918 | *0.8491* |
| *Sarcasm* | 0.9511 | 0.8130 | 0.8846 | *0.8441* |
| *Argument Basis* | 0.9534 | 0.8195 | 0.8932 | *0.8515* |
| Moral Sentiment \| Offence | 0.9517 | 0.8158 | 0.8862 | 0.8465 |
| Moral Sentiment \| Hate Speech | 0.9555 | 0.8167 | 0.9081 | 0.8551 |
| Hate Speech \| Offence | 0.9504 | 0.8386 | 0.8652 | 0.8512 |
| Sarcasm \| Moral Sentiment | 0.9533 | 0.8008 | 0.9093 | **0.8575** |
| Sarcasm \| Hate Speech | **0.9558** | 0.8191 | 0.9079 | 0.8566 |
| Sarcasm \| Offence | 0.9522 | 0.7957 | 0.9068 | 0.8469 |
| Argument Basis \| Hate Speech | 0.9521 | 0.8148 | 0.8889 | 0.8421 |
| Argument Basis \| Moral Sentiment | 0.9538 | 0.8101 | 0.9038 | 0.8492 |
| Argument Basis \| Sarcasm | 0.9535 | 0.8376 | 0.8808 | *0.8575* |
| Argument Basis \| Offence | 0.9542 | 0.8093 | 0.9066 | 0.8496 |
| Moral Sentiment \| Hate Speech \| Offence | 0.9529 | 0.7956 | 0.9121 | 0.8419 |
| Sarcasm \| Moral Sentiment \| Hate Speech | 0.9533 | 0.7938 | 0.917 | 0.8402 |
| Sarcasm \| Moral Sentiment \| Offence | 0.9523 | 0.7861 | **0.9184** | 0.8368 |
| Argument Basis \| Moral Sentiment \| Sarcasm | 0.9526 | 0.8214 | 0.8869 | 0.8503 |
| Argument Basis \| Moral Sentiment \| Hate Speech | 0.9516 | 0.8469 | 0.8661 | 0.8562 |
| Argument Basis \| Moral Sentiment \| Offence | 0.953 | 0.8248 | 0.8867 | 0.8523 |
| Sarcasm \| Moral Sentiment \| Hate Speech \| Offence | 0.9548 | 0.8117 | 0.9086 | 0.8519 |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Hate Speech | 0.9514 | 0.8315 | 0.8740 | 0.8511 |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Offence | 0.9517 | 0.8098 | 0.8905 | 0.8442 |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Offence \| Hate Speech | 0.9534 | 0.8039 | 0.9067 | 0.8459 |

**Table 6.14** Experimental model validation scores on the *Toxicity* dataset. Configurations written in *italic* signify the auxiliary tasks chosen for further exploration and **bolded** scores indicate the highest performances.

| Aux | Batch Size | Main Task Weight | Aux Task Weights | Dropout | Embedding Dim | Epochs | Hidden Dims | Learning Rate | Non-linearity | Optimiser | Shared Dim |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Hate Expert | 64 | 0.7 | 0.3 | 0.3705 | 200 | 50 | 200,200 | 0.9084 | ReLU | ASGD | 128 |
| Toxicity | 64 | 0.7 | 0.3 | 0.02884 | 100 | 100 | 64,64 | 0.3873 | Tanh | ASGD | 128 |
| Offence | 32 | 0.7 | 0.3 | 0.4568 | 64 | 100 | 200,200 | 0.2846 | ReLU | ASGD | 256 |
| Moral Sentiment | 64 | 0.7 | 0.3 | 0.1954 | 300 | 100 | 100,100 | 0.06402 | ReLU | ASGD | 256 |
| Sarcasm | 64 | 0.7 | 0.3 | 0.4534 | 300 | 100 | 300,300 | 0.3894 | ReLU | ASGD | 128 |
| Argument Basis | 64 | 0.7 | 0.3 | 0.1556 | 100 | 50 | 100,100 | 0.4948 | ReLU | ASGD | 256 |
| Hate Expert \| Offence | 64 | 0.6 | 0.2 | 0.3703 | 200 | 50 | 200,200,200 | 0.9429 | ReLU | ASGD | 128 |
| Hate Expert \| Moral Sentiment | 64 | 0.6 | 0.2 | 0.311 | 100 | 50 | 300,300,300 | 0.6185 | ReLU | ASGD | 128 |
| Offence \| Moral Sentiment | 16 | 0.6 | 0.2 | 0.1408 | 300 | 100 | 64,64,64 | 0.1237 | Tanh | SGD | 128 |
| Sarcasm \| Hate Expert | 64 | 0.6 | 0.2 | 0.3054 | 100 | 50 | 64,64,64 | 0.06252 | ReLU | SGD | 64 |
| Sarcasm \| Offence | 16 | 0.6 | 0.2 | 0.4576 | 300 | 200 | 100,100,100 | 0.2276 | ReLU | ASGD | 256 |
| Sarcasm \| Moral Sentiment | 64 | 0.6 | 0.2 | 0.3586 | 300 | 50 | 100,100,100 | 0.3822 | ReLU | ASGD | 64 |
| Hate Expert \| Offence \| Moral Sentiment | 16 | 0.6 | 0.133333333 | 0.414 | 200 | 100 | 300,300,300,300 | 0.8435 | ReLU | ASGD | 256 |
| Sarcasm \| Hate Expert \| Moral Sentiment | 32 | 0.6 | 0.133333333 | 0.152 | 64 | 100 | 200,200,200,200 | 0.3459 | ReLU | ASGD | 256 |
| Sarcasm \| Offence \| Moral Sentiment | 64 | 0.6 | 0.133333333 | 0.05853 | 300 | 100 | 64,64,64,64 | 0.04528 | ReLU | ASGD | 64 |
| Sarcasm \| Hate Expert \| Offence \| Moral Sentiment | 16 | 0.6 | 0.1 | 0.143 | 64 | 100 | 64,64,64,64,64 | 0.2368 | ReLU | ASGD | 256 |

**Table 6.15** Hyper-parameters for best performing MTL models optimised on the *Hate Speech* dataset.

| Aux | Batch Size | Main Task Weight | Aux Task Weights | Dropout | Embedding Dim | Epochs | Hidden Dims | Learning Rate | Non-linearity | Optimiser | Shared Dim |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Hate Expert | 32 | 0.7 | 0.3 | 0.06554 | 100 | 200 | 300,300 | 0.2921 | ReLU | ASGD | 256 |
| Hate Speech | 64 | 0.7 | 0.3 | 0.1785 | 100 | 50 | 100,100 | 0.281 | ReLU | SGD | 128 |
| Toxicity | 64 | 0.7 | 0.3 | 0.4405 | 300 | 50 | 200,200 | 0.3469 | ReLU | ASGD | 256 |
| Sarcasm | 32 | 0.7 | 0.3 | 0.1476 | 300 | 50 | 200,200 | 0.9616 | ReLU | ASGD | 128 |
| Argument Basis | 16 | 0.7 | 0.3 | 0.2952 | 300 | 100 | 64,64 | 0.4306 | ReLU | ASGD | 64 |
| Moral Sentiment | 32 | 0.7 | 0.3 | 0.1953 | 200 | 100 | 300,300 | 0.1415 | ReLU | SGD | 64 |
| Hate Speech \| Toxicity | 64 | 0.6 | 0.2 | 0.1663 | 100 | 50 | 100,100,100 | 0.3764 | ReLU | SGD | 256 |
| Sarcasm \| Hate Speech | 32 | 0.6 | 0.2 | 0.1497 | 100 | 100 | 300,300,300 | 0.2229 | ReLU | ASGD | 256 |
| Sarcasm \| Toxicity | 32 | 0.6 | 0.2 | 0.166 | 64 | 100 | 300,300,300 | 0.511 | ReLU | ASGD | 64 |
| Argument Basis \| Hate Speech | 64 | 0.6 | 0.2 | 0.0265 | 100 | 200 | 200,200,200 | 0.4188 | ReLU | ASGD | 256 |
| Argument Basis \| Toxicity | 64 | 0.6 | 0.2 | 0.3497 | 300 | 100 | 200,200,200 | 0.3466 | ReLU | ASGD | 128 |
| Argument Basis \| Sarcasm | 64 | 0.6 | 0.2 | 0.4527 | 200 | 100 | 64,64,64 | 0.509 | ReLU | ASGD | 256 |
| Sarcasm \| Toxicity \| Hate Speech | 64 | 0.6 | 0.133333333 | 0.4113 | 300 | 100 | 200,200,200,200 | 0.1113 | ReLU | ASGD | 256 |
| Argument Basis \| Hate Speech \| Toxicity | 16 | 0.6 | 0.133333333 | 0.2439 | 200 | 200 | 100,100,100,100 | 0.8852 | ReLU | ASGD | 64 |
| Argument Basis \| Hate Speech \| Sarcasm | 64 | 0.6 | 0.133333333 | 0.3725 | 200 | 200 | 200,200,200,200 | 0.3176 | ReLU | ASGD | 64 |
| Argument Basis \| Toxicity \| Sarcasm | 64 | 0.6 | 0.133333333 | 0.259 | 64 | 100 | 300,300,300,300 | 0.6679 | ReLU | ASGD | 128 |
| Argument Basis \| Hate Speech \| Toxicity \| Sarcasm | 64 | 0.6 | 0.1 | 0.0103 | 64 | 100 | 64,64,64,64,64 | 0.4785 | ReLU | ASGD | 64 |

**Table 6.16** Hyper-parameters for best performing MTL models optimised on the *Offence* dataset.

| Aux | Batch Size | Main Task Weight | Aux Task Weights | Dropout | Embedding Dim | Epochs | Hidden Dims | Learning Rate | Non-linearity | Optimiser | Shared Dim |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Waseem | 64 | 0.7 | 0.3 | 0.1166 | 300 | 200 | 64,64 | 0.4055 | ReLU | ASGD | 256 |
| Hate Speech | 64 | 0.7 | 0.3 | 0.3346 | 100 | 50 | 300,300 | 0.003051 | ReLU | AdamW | 256 |
| Offence | 64 | 0.7 | 0.3 | 0.03044 | 300 | 200 | 200,200 | 0.3432 | ReLU | SGD | 128 |
| Moral Sentiment | 64 | 0.7 | 0.3 | 0.3558 | 300 | 200 | 64,64 | 0.7853 | ReLU | ASGD | 256 |
| Sarcasm | 64 | 0.7 | 0.3 | 0.3711 | 200 | 200 | 100,100 | 0.9143 | ReLU | ASGD | 256 |
| Argument Basis | 64 | 0.7 | 0.3 | 0.3995 | 300 | 50 | 200,200 | 0.5104 | ReLU | SGD | 128 |
| Hate Speech \| Offence | 64 | 0.6 | 0.2 | 0.03108 | 300 | 200 | 100,100,100 | 0.2465 | ReLU | SGD | 256 |
| Moral Sentiment \| Hate Speech | 64 | 0.6 | 0.2 | 0.3406 | 300 | 200 | 200,200,200 | 0.2291 | ReLU | SGD | 64 |
| Moral Sentiment \| Offence | 64 | 0.6 | 0.2 | 0.08389 | 200 | 200 | 64,64,64 | 0.8606 | ReLU | ASGD | 128 |
| Sarcasm \| Moral Sentiment | 64 | 0.6 | 0.2 | 0.2051 | 200 | 200 | 300,300,300 | 0.8602 | ReLU | ASGD | 64 |
| Sarcasm \| Hate Speech | 64 | 0.6 | 0.2 | 0.2315 | 300 | 200 | 300,300,300 | 0.2219 | ReLU | SGD | 128 |
| Sarcasm \| Offence | 64 | 0.6 | 0.2 | 0.02304 | 100 | 200 | 64,64,64 | 0.804 | ReLU | ASGD | 256 |
| Moral Sentiment \| Hate Speech \| Offence | 64 | 0.6 | 0.133333333 | 0.3018 | 200 | 200 | 300,300,300,300 | 0.9543 | ReLU | ASGD | 256 |
| Sarcasm \| Moral Sentiment \| Hate Speech | 64 | 0.6 | 0.133333333 | 0.2762 | 200 | 100 | 300,300,300,300 | 0.4007 | ReLU | SGD | 128 |
| Sarcasm \| Moral Sentiment \| Offence | 64 | 0.6 | 0.133333333 | 0.2939 | 64 | 200 | 200,200,200,200 | 0.8591 | ReLU | ASGD | 64 |
| Sarcasm \| Moral Sentiment \| Hate Speech \| Offence | 32 | 0.6 | 0.1 | 0.1936 | 300 | 50 | 100,100,100,100,100 | 0.004907 | ReLU | AdamW | 64 |
| Argument Basis \| Moral Sentiment | 64 | 0.6 | 0.2 | 0.01288 | 200 | 200 | 300,300,300 | 0.968 | ReLU | ASGD | 256 |
| Argument Basis \| Sarcasm | 64 | 0.6 | 0.2 | 0.061 | 200 | 200 | 300,300,300 | 0.384 | ReLU | SGD | 256 |
| Argument Basis \| Hate Speech | 64 | 0.6 | 0.2 | 0.09094 | 200 | 200 | 300,300,300 | 0.8456 | ReLU | ASGD | 128 |
| Argument Basis \| Offence | 64 | 0.6 | 0.2 | 0.2733 | 300 | 100 | 100,100,100 | 0.3757 | ReLU | SGD | 256 |
| Argument Basis \| Moral Sentiment \| Sarcasm | 64 | 0.6 | 0.133333333 | 0.03792 | 300 | 200 | 300,300,300,300 | 0.934 | ReLU | ASGD | 64 |
| Argument Basis \| Moral Sentiment \| Hate Speech | 64 | 0.6 | 0.133333333 | 0.05325 | 300 | 200 | 300,300,300,300 | 0.4349 | ReLU | SGD | 64 |
| Argument Basis \| Moral Sentiment \| Offence | 64 | 0.6 | 0.133333333 | 0.1082 | 300 | 200 | 100,100,100,100 | 0.3236 | ReLU | SGD | 128 |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Hate Speech | 64 | 0.6 | 0.1 | 0.06104 | 100 | 200 | 300,300,300,300,300 | 0.565 | ReLU | SGD | 256 |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Offence | 64 | 0.6 | 0.1 | 0.2149 | 300 | 200 | 100,100,100,100,100 | 0.6336 | ReLU | SGD | 128 |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Offence \| Hate Speech | 64 | 0.6 | 0.08 | 0.2348 | 100 | 100 | 200,200,200,200,200,200 | 0.546 | ReLU | SGD | 256 |

**Table 6.17** Hyper-parameters for best performing MTL models optimised on the *Toxicity* dataset.

the *Hate Expert* dataset outperform all baseline models on the validation set in terms of F1-score. Thus, all auxiliary tasks, asides from the *Hate Expert* task, are selected for further experimentation.

#### 6.3.2.2 Evaluation Data Performances

Turning to the performance on the test sets, the results presented in tables 6.18 to 6.20 are the means of five runs with different random seeds and the standard deviation for each metric.

**Hate Speech Main Task** Considering first the results in table 6.18 for the MTL models optimised for the *Hate Speech* task. All but one dataset configurations yield an improved accuracy over the MLP baseline, though none outperform the SVM baseline. The best auxiliary task configuration for accuracy is one that uses the *Hate Expert* and the *Sarcasm* auxiliary tasks. Similarly to the results on the validation set, most MTL MLP models yield an improvement in terms of recall over all baselines, while only one shows a strong decrease in performance. The best auxiliary task configuration in terms of recall only uses the *Toxicity* dataset for an auxiliary task. In terms of precision, all task configurations provide an improvement over the MLP baseline, though similarly to the case with the accuracy score, none outperform the best-performing baseline in terms of precision, the linear SVM. Although all task configurations yield an improvement over the MLP baseline, an interesting issue occurs where the use of all auxiliary tasks has a detrimental effect, resulting in the worst performance of the MTL models in terms of precision. For the F1-score, no models outperform the SVM baseline, though all models outperform the MLP baseline. The strongest performance here is obtained by an auxiliary task combination where two out of three auxiliary tasks are non-abusive in nature. This performance gain is obtained through an increase in the experimental model's ability in precision at the cost of a slightly lower recall score. Although, on their own, none of the non-abusive tasks obtain the highest performances out of all auxiliary task configurations, they frequently post highly competitive scores with the all other configurations. This suggests that some of the improvements obtained when using multiple auxiliary tasks is obtained through the use of non-abusive information encoded into the shared layer of the model.

Aligning with the insights on dataset mappings identified in chapter 5, the *Hate Expert* dataset provides for boosted predicted power when used in conjunction with a non-abusive auxiliary task. Somewhat surprisingly, the *Offence* auxiliary task equally provides boosts in performance. These improvements in performance indicate that the source domain of the

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Hate Expert | 0.8082 ($\sigma$ 0.0098) | 0.7496 ($\sigma$ 0.0117) | 0.7661 ($\sigma$ 0.0153) | 0.7564 ($\sigma$ 0.0044) |
| Offence | 0.8194 ($\sigma$ 0.0047) | 0.7001 ($\sigma$ 0.0181) | **0.8204** ($\sigma$ 0.0069) | 0.7430 ($\sigma$ 0.0124) |
| Toxicity | 0.8105 ($\sigma$ 0.0075) | 0.7236 ($\sigma$ 0.0244) | 0.7779 ($\sigma$ 0.0184) | 0.7424 ($\sigma$ 0.0165) |
| Moral Sentiment | 0.8234 ($\sigma$ 0.0024) | 0.7302 ($\sigma$ 0.0120) | 0.8039 ($\sigma$ 0.0079) | 0.7595 ($\sigma$ 0.0067) |
| Argument Basis | 0.8141 ($\sigma$ 0.0090) | 0.7380 ($\sigma$ 0.0185) | 0.7744 ($\sigma$ 0.0228) | 0.7489 ($\sigma$ 0.0146) |
| Sarcasm | 0.8192 ($\sigma$ 0.0018) | 0.7262 ($\sigma$ 0.0167) | 0.7993 ($\sigma$ 0.0145) | 0.7548 ($\sigma$ 0.0065) |
| Hate Expert \| Moral Sentiment | 0.8173 ($\sigma$ 0.0071) | 0.7366 ($\sigma$ 0.0089) | 0.7882 ($\sigma$ 0.0191) | 0.7523 ($\sigma$ 0.0100) |
| Offence \| Moral Sentiment | 0.7985 ($\sigma$ 0.0090) | 0.6994 ($\sigma$ 0.0240) | 0.7628 ($\sigma$ 0.0155) | 0.7244 ($\sigma$ 0.0163) |
| Hate Expert \| Offence | 0.8147 ($\sigma$ 0.0053) | 0.7294 ($\sigma$ 0.0222) | 0.7845 ($\sigma$ 0.0184) | 0.7505 ($\sigma$ 0.0070) |
| Sarcasm \| Hate Expert | **0.8253** ($\sigma$ 0.0079) | 0.7333 ($\sigma$ 0.0277) | 0.8096 ($\sigma$ 0.0167) | 0.7618 ($\sigma$ 0.0171) |
| Sarcasm \| Offence | 0.8182 ($\sigma$ 0.0051) | 0.7186 ($\sigma$ 0.0266) | 0.8031 ($\sigma$ 0.0139) | 0.7518 ($\sigma$ 0.0160) |
| Sarcasm \| Moral Sentiment | 0.8215 ($\sigma$ 0.0071) | 0.7209 ($\sigma$ 0.0231) | 0.8108 ($\sigma$ 0.0101) | 0.7548 ($\sigma$ 0.0164) |
| Hate Expert \| Offence \| Moral Sentiment | 0.7901 ($\sigma$ 0.0176) | 0.7488 ($\sigma$ 0.0057) | 0.7366 ($\sigma$ 0.0227) | 0.7407 ($\sigma$ 0.0153) |
| Sarcasm \| Hate Expert \| Moral Sentiment | 0.8180 ($\sigma$ 0.0071) | 0.7129 ($\sigma$ 0.0222) | 0.8016 ($\sigma$ 0.0046) | 0.7473 ($\sigma$ 0.0166) |
| Sarcasm \| Offence \| Moral Sentiment | 0.8151 ($\sigma$ 0.0066) | **0.7502** ($\sigma$ 0.0070) | 0.7762 ($\sigma$ 0.0103) | **0.7623** ($\sigma$ 0.0051) |
| Sarcasm \| Hate Expert \| Offence \| Moral Sentiment | 0.8124 ($\sigma$ 0.0101) | 0.6898 ($\sigma$ 0.0221) | 0.8142 ($\sigma$ 0.0053) | 0.7278 ($\sigma$ 0.0234) |

**Table 6.18** Experimental model evaluation scores on the *Hate Speech* dataset. **bolded** scores indicate the highest performances and $\sigma$ values indicate the standard deviation.

data offer greater inductive biases, that can be optimised, than the size, as evidenced by the *Toxicity* dataset not being selected for further experimentation. The selection of the *Hate Expert* auxiliary task also suggests that an alignment of dataset goals, and the processes to achieve those goals, also have a positive factor.

**Offence Main Task** When using *Offence* detection as the main task, most models outperform all baselines in terms of `F1-score` (see table 6.19). The exception to this pattern are two auxiliary task configurations: the first where only the *Argument Basis* auxiliary task is used, and the second where the *Sarcasm* auxiliary task is used in conjunction with one abusive language detection task. In the setting where only one auxiliary task is used, the *Hate Speech* auxiliary task somewhat surprisingly provides for slightly bigger improvements across all scores than the *Toxicity* auxiliary task. This further evidences that abusive tasks from the same data source provide for a good source of data. When combining multiple auxiliary tasks however, the *Toxicity* data provides for more stable improvements across different auxiliary task combinations. This further lends credibility to the observation that a similarity of dataset goals is useful in a multi-task setting.

Focusing on the non-abusive auxiliary tasks, the *Sarcasm* and *Argument Basis* auxiliary tasks are selected for further experimentation given the results on the validation data. The *Sarcasm* auxiliary task, when used in isolation from other auxiliary tasks, offers competitive results with the abusive language detection auxiliary tasks used in isolation, and outperforms the closely related *Toxicity* auxiliary task in terms of `accuracy`. The *Argument Basis* auxiliary task affords performance improvements along `accuracy`, `precision`, and `F1-score` and only when this task is used in conjunction with other auxiliary tasks, specifically the *Sarcasm*

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Hate Speech | 0.8970 ($\sigma$ 0.0028) | 0.6952 ($\sigma$ 0.0240) | 0.7685 ($\sigma$ 0.0339) | 0.7119 ($\sigma$ 0.0138) |
| Toxicity | 0.8916 ($\sigma$ 0.0036) | 0.6829 ($\sigma$ 0.0176) | 0.7659 ($\sigma$ 0.0074) | 0.7099 ($\sigma$ 0.0159) |
| Hate Expert | 0.8871 ($\sigma$ 0.0048) | 0.6794 ($\sigma$ 0.0274) | 0.7552 ($\sigma$ 0.0110) | 0.7028 ($\sigma$ 0.0193) |
| Sarcasm | 0.8933 ($\sigma$ 0.0029) | 0.6790 ($\sigma$ 0.0303) | 0.7594 ($\sigma$ 0.0141) | 0.6987 ($\sigma$ 0.0271) |
| Argument Basis | 0.8973 ($\sigma$ 0.0045) | 0.6740 ($\sigma$ 0.0238) | 0.7775 ($\sigma$ 0.0202) | 0.6847 ($\sigma$ 0.0207) |
| Moral Sentiment | 0.8956 ($\sigma$ 0.0045) | 0.6860 ($\sigma$ 0.0344) | 0.7541 ($\sigma$ 0.0267) | 0.6938 ($\sigma$ 0.0356) |
| Hate Speech \| Toxicity | 0.8860 ($\sigma$ 0.0046) | 0.6961 ($\sigma$ 0.0157) | 0.7294 ($\sigma$ 0.0211) | 0.7033 ($\sigma$ 0.0135) |
| Sarcasm \| Hate Speech | 0.8891 ($\sigma$ 0.0080) | 0.6561 ($\sigma$ 0.0239) | 0.7757 ($\sigma$ 0.0157) | 0.6806 ($\sigma$ 0.0200) |
| Sarcasm \| Toxicity | 0.8977 ($\sigma$ 0.0011) | 0.6723 ($\sigma$ 0.0126) | **0.7899** ($\sigma$ 0.0063) | 0.6811 ($\sigma$ 0.0227) |
| Sarcasm \| Toxicity \| Hate Speech | 0.8923 ($\sigma$ 0.0028) | 0.6877 ($\sigma$ 0.0213) | 0.7601 ($\sigma$ 0.0079) | 0.7091 ($\sigma$ 0.0151) |
| Argument Basis \| Hate Speech | 0.8864 ($\sigma$ 0.0036) | 0.7043 ($\sigma$ 0.0320) | 0.7386 ($\sigma$ 0.0173) | 0.7155 ($\sigma$ 0.0139) |
| Argument Basis \| Toxicity | 0.8917 ($\sigma$ 0.0052) | 0.6929 ($\sigma$ 0.0203) | 0.7552 ($\sigma$ 0.0049) | 0.7143 ($\sigma$ 0.0139) |
| Argument Basis \| Sarcasm | 0.9010 ($\sigma$ 0.0022) | 0.6939 ($\sigma$ 0.0170) | 0.7791 ($\sigma$ 0.0110) | 0.7105 ($\sigma$ 0.0190) |
| Argument Basis \| Hate Speech \| Toxicity | 0.8972 ($\sigma$ 0.0036) | 0.6861 ($\sigma$ 0.0089) | 0.7723 ($\sigma$ 0.0085) | 0.7064 ($\sigma$ 0.0083) |
| Argument Basis \| Hate Speech \| Sarcasm | **0.9025 ($\sigma$ 0.0007)** | **0.7107 ($\sigma$ 0.0090)** | 0.7820 ($\sigma$ 0.0093) | **0.7291 ($\sigma$ 0.0072)** |
| Argument Basis \| Toxicity \| Sarcasm | 0.8925 ($\sigma$ 0.0044) | 0.6973 ($\sigma$ 0.0218) | 0.7472 ($\sigma$ 0.0082) | 0.7138 ($\sigma$ 0.0125) |
| Argument Basis \| Hate Speech \| Toxicity \| Sarcasm | 0.8948 ($\sigma$ 0.0016) | 0.6979 ($\sigma$ 0.0048) | 0.7652 ($\sigma$ 0.0102) | 0.7100 ($\sigma$ 0.0049) |

**Table 6.19** Experimental model evaluation scores on the *Offence* dataset. **bolded** scores indicate the highest performances and $\sigma$ values indicate the standard deviation.

and *Hate Speech* auxiliary tasks, do the models outperform the best baseline model in terms of `precision`.

The experimental models optimised for the *Offence* detection task post a poorer `recall` than the best performing baseline model, thus the increased performances in `F1-score` are driven by improvements in `precision` score. The best performing baseline model, in terms of `recall`, obtains a score of 0.91 but only obtains 0.57 in terms of `precision`. Thus, adding auxiliary tasks provides for models that are more balanced in terms of performance on `precision` and `recall`, yielding an improved performance in terms of `F1-score` as neither `precision` or `recall` are being neglected in favour of the other. Comparing the MTL models to the SVM baseline, all auxiliary task settings models strongly outperform the baseline in terms of `recall`.

**Toxicity Main Task** The final task that I explore experimentally through MTL is the *Toxicity* detection task. This task is the task that has the most auxiliary tasks that are explored experimentally with five out of six tasks explored, the only task that is note explored is the *Hate Expert* task. Moreover, this is also the task with the largest dataset available and the only binary task, meaning that this is the only task where the results are directly comparable with the results obtained in chapter 5.

Observing experimental model results in table 6.20 with the performances of the baseline models in table 6.11, we see that no setting of the auxiliary tasks outperform the SVM baseline in terms of `F1-score` or `accuracy`. However, all task configurations outperform the SVM in terms of `recall` and are competitive in terms of the `accuracy` score with the

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Hate Speech | 0.9523 ($\sigma$ 0.0006) | 0.8265 ($\sigma$ 0.0184) | 0.8836 ($\sigma$ 0.01505) | 0.8513 ($\sigma$ 0.0052) |
| Offence | 0.9506 ($\sigma$ 0.0027) | **0.8445 ($\sigma$ 0.0091)** | 0.8637 ($\sigma$ 0.0147) | 0.8535 ($\sigma$ 0.0049) |
| Hate Expert | 0.9497 ($\sigma$ 0.0008) | 0.7794 ($\sigma$ 0.0081) | 0.9074 ($\sigma$ 0.0124) | 0.8283 ($\sigma$ 0.0027) |
| Moral Sentiment | 0.9535 ($\sigma$ 0.0005) | 0.8058 ($\sigma$ 0.0074) | 0.9061 ($\sigma$ 0.0088) | 0.8469 ($\sigma$ 0.0025) |
| Sarcasm | 0.9532 ($\sigma$ 0.0004) | 0.8029 ($\sigma$ 0.0131) | 0.9078 ($\sigma$ 0.0130) | 0.8452 ($\sigma$ 0.0050) |
| Argument Basis | 0.9477 ($\sigma$ 0.0040) | 0.7604 ($\sigma$ 0.0307) | 0.9152 ($\sigma$ 0.0098) | 0.8145 ($\sigma$ 0.0245) |
| Moral Sentiment \| Offence | 0.9518 ($\sigma$ 0.0005) | 0.7897 ($\sigma$ 0.0125) | 0.9118 ($\sigma$ 0.0144) | 0.8370 ($\sigma$ 0.0048) |
| Moral Sentiment \| Hate Speech | 0.9530 ($\sigma$ 0.0013) | 0.7984 ($\sigma$ 0.0163) | 0.9109 ($\sigma$ 0.0114) | 0.8430 ($\sigma$ 0.0088) |
| Hate Speech \| Offence | **0.9538 ($\sigma$ 0.0008)** | 0.8238 ($\sigma$ 0.0142) | 0.8931 ($\sigma$ 0.0139) | **0.8537 ($\sigma$ 0.0038)** |
| Sarcasm \| Moral Sentiment | 0.9526 ($\sigma$ 0.0007) | 0.7929 ($\sigma$ 0.0092) | 0.9132 ($\sigma$ 0.0054) | 0.8401 ($\sigma$ 0.0052) |
| Sarcasm \| Hate Speech | 0.8947 ($\sigma$ 0.0798) | 0.7800 ($\sigma$ 0.0733) | 0.8467 ($\sigma$ 0.0535) | 0.8070 ($\sigma$ 0.0654) |
| Sarcasm \| Offence | 0.9523 ($\sigma$ 0.0007) | 0.7949 ($\sigma$ 0.0138) | 0.9099 ($\sigma$ 0.0113) | 0.8402 ($\sigma$ 0.0065) |
| Argument Basis \| Hate Speech | 0.9514 ($\sigma$ 0.0008) | 0.8027 ($\sigma$ 0.0153) | 0.8966 ($\sigma$ 0.0159) | 0.8410 ($\sigma$ 0.0057) |
| Argument Basis \| Moral Sentiment | 0.9528 ($\sigma$ 0.0004) | 0.8026 ($\sigma$ 0.0070) | 0.9047 ($\sigma$ 0.0060) | 0.8443 ($\sigma$ 0.0031) |
| Argument Basis \| Sarcasm | 0.9528 ($\sigma$ 0.0011) | 0.8182 ($\sigma$ 0.0224) | 0.8925 ($\sigma$ 0.0144) | 0.8495 ($\sigma$ 0.0095) |
| Argument Basis \| Offence | 0.9511 ($\sigma$ 0.0014) | 0.7813 ($\sigma$ 0.0132) | 0.9162 ($\sigma$ 0.0139) | 0.8321 ($\sigma$ 0.0079) |
| Moral Sentiment \| Hate Speech \| Offence | 0.9532 ($\sigma$ 0.0005) | 0.7932 ($\sigma$ 0.0095) | 0.9174 ($\sigma$ 0.0100) | 0.8415 ($\sigma$ 0.0042) |
| Sarcasm \| Moral Sentiment \| Hate Speech | 0.9517 ($\sigma$ 0.0012) | 0.7845 ($\sigma$ 0.0115) | 0.9165 ($\sigma$ 0.0051) | 0.8348 ($\sigma$ 0.0072) |
| Sarcasm \| Moral Sentiment \| Offence | 0.9513 ($\sigma$ 0.0012) | 0.7760 ($\sigma$ 0.0106) | **0.9235 ($\sigma$ 0.0065)** | 0.8303 ($\sigma$ 0.0071) |
| Argument Basis \| Moral Sentiment \| Sarcasm | 0.9535 ($\sigma$ 0.0005) | 0.8053 ($\sigma$ 0.0160) | 0.9083 ($\sigma$ 0.0177) | 0.8467 ($\sigma$ 0.0054) |
| Argument Basis \| Moral Sentiment \| Hate Speech | 0.9532 ($\sigma$ 0.0018) | 0.8170 ($\sigma$ 0.0075) | 0.8947 ($\sigma$ 0.0133) | 0.8503 ($\sigma$ 0.0046) |
| Argument Basis \| Moral Sentiment \| Offence | 0.9537 ($\sigma$ 0.0007) | 0.8194 ($\sigma$ 0.0142) | 0.8958 ($\sigma$ 0.0102) | 0.8520 ($\sigma$ 0.0057) |
| Sarcasm \| Moral Sentiment \| Hate Speech \| Offence | 0.9526 ($\sigma$ 0.0018) | 0.7883 ($\sigma$ 0.0228) | 0.9197 ($\sigma$ 0.0131) | 0.8379 ($\sigma$ 0.0128) |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Hate Speech | 0.9526 ($\sigma$ 0.0003) | 0.8242 ($\sigma$ 0.0190) | 0.8869 ($\sigma$ 0.0155) | 0.8512 ($\sigma$ 0.0060) |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Offence | 0.9505 ($\sigma$ 0.0017) | 0.8107 ($\sigma$ 0.0232) | 0.8852 ($\sigma$ 0.0183) | 0.8417 ($\sigma$ 0.0098) |
| Argument Basis \| Moral Sentiment \| Sarcasm \| Offence \| Hate Speech | 0.9512 ($\sigma$ 0.0018) | 0.7948 ($\sigma$ 0.0236) | 0.9035 ($\sigma$ 0.0187) | 0.8374 ($\sigma$ 0.0118) |

**Table 6.20** Experimental model evaluation scores on the *Toxicity* dataset. **bolded** scores indicate the highest performances and $\sigma$ values indicate the standard deviation.

worst performing configuration, one that uses the *Sarcasm* and the *Hate Speech* auxiliary tasks scoring 0.0635 below the SVM baseline and 0.045 lower the MLP baseline. On the other hand, the best performing model in terms of `accuracy` has a score that is 0.0044 lower than the SVM baseline and 0.0141 higher than the MLP baseline. For `precision`, similarly none of the experimental settings outperform the SVM or the ensemble baselines. The best auxiliary task configuration only uses the *Offence* auxiliary task and obtains a `precision` score of 0.8445, 0.0563 lower than the SVM baseline, while the worst performing auxiliary task setting only uses the *Argument Basis* with a score of 0.7604 or 0.1404 lower than the SVM baseline. Here is an indication that using the `F1-score` as the only metric by which the auxiliary tasks are chosen is suboptimal as the *Hate Expert* auxiliary task outperforms the *Argument Basis* auxiliary task for all metrics asides from the `recall` score. Although the performance in terms of `precision` is lower than the SVM baseline, all auxiliary task configurations outperform the MLP baseline with the worst performing task setting outperforming the MLP baseline by 0.0625 and a 0.1466 improvement over the baseline for the best-performing task setting. For the `recall` score, the best performing baseline model is the ensemble baseline, closely followed by the MLP baseline. Neither of these baselines are outperformed by the MTL models though all experimental models improve on the SVM baseline. The best performing model here uses three different auxiliary tasks, namely the *Sarcasm* task, the *Moral Sentiment* task, and the *Offence* task and only underperforms by 0.0124 in comparison to the MLP baseline. Finally, focusing on the `F1-score`. Although no auxiliary task setting outperforms the best performing baseline, the SVM baseline, task configurations strongly outperform the MLP baseline. The best performing auxiliary task setting uses the *Hate Speech* and the *Offence* auxiliary tasks, that is all of the abusive language detection tasks under consideration and it outperforms the MLP baseline by 0.0905. The worst performing auxiliary task setting outscores the MLP baseline by 0.0438 and uses the *Sarcasm* and the *hate Speech* auxiliary tasks.

In relation to the classification scores achieved by the best model for the *Toxicity* dataset in chapter 5, the best performing model selected by highest `F1-score` is a LSTM model that uses word tokens as its input obtains 0.9510 in `accuracy`, 0.8047 in `precision`, 0.9058 in `recall`, and 0.8357 in `F1-score`.[4] Thus, the model from chapter 5 is outperformed, in terms of `accuracy`, `precision` and `F1-score`, as the best performing auxiliary task settings, ranked by `F1-score`. This further demonstrating the space for MTL models to improve performance on single-task models.

[4] The scores here are the mean score over five different configurations of the random seeds.

**Auxiliary Task Patterns** When considering the performances of each main task in isolation, the larger patterns across the different main tasks are obscured by the details of the model performances. Here I take a birds-eye view on all three main tasks and the patterns that emerge from across the different main tasks and their auxiliary task settings.

In the performances of all three main tasks, auxiliary task configurations that include abusive language detection tasks tend to post the high performances. This may be because forms and phrasings in abusive, and non-abusive, text that are infrequently occurring in the main task can be aided by the data provided in the abusive auxiliary tasks. This is particularly evidenced through the improvements along the `precision` score, meaning that using abusive language detection auxiliary tasks can minimise the number of false negatives that the model predicts. These increases in the `precision` scores also positively affect the `F1-scores`, although these are often hampered by a lower `recall` score. The lower performance in terms of *recall* is somewhat surprising, as one might imagine that the inclusion of more abusive data would often mean that there is an improvement along the lines of *recall*. However, as I retain the class distribution in the different splits of the dataset and shuffle the order of batches between each epoch, it may be that the same batches are frequently chosen and that when distinct batches are chosen, they do not, on aggregate, provide for enough new examples of abuse to provide a boost in terms of *recall*. Moreover, the weighting of each auxiliary task may provide an additional cause for this pattern, as each auxiliary task has an equal probability of being selected in a given configuration and not enough abusive samples are provided to the model to improve *recall*. Finally, this may also be due to using several different forms of abuse with distinct definitions of what constitutes abuse. Such disparate definitions and forms of abuse may provide competing signals to models on documents with a high degree of textual similarity. While there appears to be a lesser impact of dataset size, likely because the number of epochs is fixed, so the larger datasets are not afforded the ability to contribute more by virtue of their size, the data source and dataset goals appear to influence the performances. This is apparent in all three main task settings but is particularly visible when the main task is *Offence* detection. Here, *Hate Speech* shares the same data source but has different annotation goals whereas *Toxicity* shares in the dataset goals but the data has a different source. In the setting where only one auxiliary task is used, both post comparative performances. When two auxiliary tasks are used and these auxiliary tasks are used together, they produce results that are lower across all metrics asides from `recall` than when only one of the tasks are used. When both auxiliary abuse detection tasks are used, then adding a at least one more auxiliary task that is not abuse detection has a slightly beneficial impact on the model performance.

The non-abusive auxiliary tasks may at first seem to contribute less towards improved model performances than the abuse detection auxiliary tasks, however for two out of three different main tasks, the best `F1-scores` are obtained by the combination of abuse detection tasks and auxiliary tasks that are not addressing the question of abuse. The non-abuse related auxiliary tasks, when used in conjunction with abuse detection tasks offer modelling improvements across all metrics. However, some non-abusive tasks appear to be more applicable in general, specifically the *Sarcasm* auxiliary task appears across all three main tasks while the *Argument Basis* and *Moral Sentiment* auxiliary tasks appear for two different main task settings. Moreover, all three auxiliary tasks appear in the best performing auxiliary task configurations and in the setting where only one auxiliary task is used, they achieve competitive performances with the models that use an abuse detection auxiliary task.

## 6.4 Conclusions

Automated detection of abuse occurring in online spaces is a complex problem that is contingent on the ability to contextualise comments made by authors with the intentions of the authors, how the comments are perceived by readers, and situating the comments within the contexts of the authors. In this chapter, I attempt to address the issue of contextualisation through the use of MTL. Using MTL, I examine how different auxiliary tasks impact in-domain classification of abuse detection. Specifically, I examine how the use of datasets developed for the purpose of optimising machine learning models for abuse as auxiliary tasks can impact the in-domain performances of different forms of abuse detection. I contrast the use of these resources with resources that are developed for classifying the basis of an argument, the moral sentiments displayed in messages, and the use of sarcasm and examine how these impact the ability of MTL models to classify abuse, when the resources are used as auxiliary tasks. Finally, I examine whether there are synergies between datasets developed for detecting abuse and datasets developed to predict other constructs that are conducive towards improved in-domain classification performance for the main task.

Through the use of MTL, I show that machine learning models developed for detecting abuse can benefit from using auxiliary task datasets. In studying how other abuse detection datasets impact performance on the main task when they are used as auxiliary tasks, I find that datasets that 1) share in the goals of the main task dataset, or 2) are sampled from the same data source, even with disparate dataset goals positively influence the performance of neural machine learning models on the main task. The former condition further provides support for the findings in chapter 5 where I identified that common dataset goals allow for better generalisation. Rather than provide for better generalisation from one task to another, I find in this chapter that different datasets with similar goals can provide be used in a multi-task setting to improve the in-domain classification results for one another. The latter observation can be understood through the aims and purposes of the MTL framework. MTL operates with the express notion that distinct tasks may share inherent latent information that can be represented in a machine learning model through joint optimisation of the tasks. Where language production on online platforms is created under restrictions put in place by the platforms and the cultures of communication that are fostered on the platforms. Thus, through using auxiliary tasks for abuse where the data is sourced form the same platform(s), an MTL model can optimise representations of the particularities of communication on the platform(s) in question, thus gaining a representation that can yield improvements in the performances of the main task.

In addressing how data developed for tasks that are not directly related to abuse detection can impact the performance of abuse detection models, I find that such tasks can aid in the classification capabilities of abuse detection models, when the choice of tasks is informed by specific questions surrounding the primary task. In particular, I find that the tasks of identifying whether an argument is made on an emotional or factual basis, detecting sarcasm, and detecting the moral sentiments all have a beneficial impact across the different main task settings. While I find that all non-abusive auxiliary tasks that I experiment with are beneficial for some forms of abuse, I also find that not all such tasks are equally well suited for all forms of abuse. For instance, the *Argument Basis* auxiliary task does not perform well enough to be included in the experiments, when it is used as the only auxiliary task to the *Hate Speech* main task. Similarly, when the main task is the *Offence* task, the *Moral Sentiment* auxiliary task does not perform well enough on the validation data to warrant inclusion in the experiments.

Finally, considering how the abusive and non-abusive auxiliary tasks interact when used in conjunction with each other, the results obtained in this chapter are clear. I find that there are clear benefits to the main task from using a combination of abusive and non-abusive auxiliary tasks with several task configurations that include abusive and non-abusive auxiliary tasks posting highly competitive scores with other top performing models, if not outperforming all other task configurations outright. By identifying abusive and non-abusive auxiliary tasks that perform well in the case where there is only one auxiliary task, it is possible for a model that uses them can benefit from the selected task to post performances that improve on the scores achieved in the single-auxiliary task setting. Moreover, as multiple auxiliary tasks are used, MTL-based models also optimise representations of each task, thus they come to closer reflect how people speak. This is particularly clear when the auxiliary task is *Sarcasm* detection, where models optimise representations of what constitutes abuse in addition to how sarcasm is expressed in the auxiliary task data. Thus, MTL models hold the potential for more closely representing different facets of speech, and subsequently different facets of speakers, than single-task machine learning models.

### 6.4.1 Future Work

The findings in this chapter have implications in the way automated abuse detection has been performed to date. Beyond the limited number of auxiliary tasks that have been investigated in this chapter, there is space for further experimentation on different kinds of auxiliary tasks. For instance, two natural extensions of this work in terms of auxiliary task selection are apparent: 1) the investigation of how optimising for core NLP tasks can yield a potential benefit as these can aid MTL models in representing language construction in different data sources and 2) following on the work of Davidson et al. (2019) and Dias Oliva et al. (2021), using datasets that specifically incorporate varieties of English spoken by groups that are often subject to high false positive errors can allow designers to encode in the model how different groups are affected and should be affected by machine learning models for content moderations. Building on my findings of how similarity in data sources impact the main task, there is space for investigating how non-abusive tasks that are collected from the same data source can impact the modelling performance.

Beyond a consideration of different auxiliary tasks there are also several modelling questions that could be further explored. First, from my experiments I find that the models tend to prefer the largest batch size that I experiment with, thus an investigation on the impacts of larger batch sizes is warranted. The question of how to weight the different auxiliary tasks and the frequency with which batches are selected from each also warrants further investigation as all auxiliary tasks are unlikely to contribute equally, thus identifying dynamic methods for weighting each task or setting the frequency with which batches are chosen from each task may also provide an avenue for improved performances. Finally, in my experimental models I do not use pre-optimised embedding layers in my models, however the use of these have been shown to achieve high classification performances in single-task neural networks, thus the use of pre-optimised embedding layers may allow for further improvements of the classification scores achieved.

## 6.5 Summary

In this chapter, I have sought to examine Multi-Task Learning as a method to incorporate greater context into machine learning models than what is provided in only datasets for abuse. Thus, I have attempted to address *RQ II* by examining which auxiliary tasks provide for useful contextualisation for abuse detection. That is, I have sought to answer *RQ 3*: How do the individual and combinatory use of abuse classification and non-abusive tasks impact classification of specific forms of abuse?

From my experiments, I find that MTL can be a useful technique for abuse detection due to the joint representation of multiple tasks that is encoded in the model. I further find that selecting the appropriate auxiliary tasks is heavily dependent on the primary task dataset and the goals the authors had in developing the resource.

These findings make clear the potential for contemporaneous machine learning techniques to more closely embody the subjectivities of an individual. Returning briefly to the infrastructure of content moderation and content moderation as a third party reader, the proximity between the values held by the third party reader and the recipient is strongly indicative of the abilities of content moderation infrastructures to more accurately represent a given recipient. MTL for hate speech detection, and more generally abuse detection, then offers a means through which practitioners can begin to develop machine learning models that more closely represent the readers whose behalf they are purported to act on behalf of.

# Chapter 7

# State of the Art White Supremacy: On Disembodiment in the Machine Learning Pipeline[1]

> What does it mean when the tools of a racist patriarchy are used to examine the fruits of that same patriarchy? It means that only the most narrow parameters of change are possible and allowable.
>
> Lorde (1984, p.110-111)

In the previous chapters we have identified different areas of concern for the use of models and data. From the ways in which content moderation technologies come to create discriminatory outcomes in chapter 4, to the constraints of document transformation in chapter 5, and the influence of multiple data sources and prediction tasks in chapter 6. In these chapters, I sought to find means to contextualise computational methods, and to make them more closely represent the subjectivities and contexts of speakers from within frames of existing computational methods. As the chapters collectively point to, there is an inherent limitation to what is achievable within computational pipelines in which the entire process, from dataset creation to modelling, is not developed while respecting human subjectivities. More precisely, the need for finding ways to approximate contexts and subjectivities highlights how the current machine learning pipeline does not specify how subjectivities are embedded in these technologies. Understanding precisely where such shortcomings arise in

[1] This chapter contains elements from a collaboration with Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. The associated paper is currently under review in the Conference on Fairness, Accountability, and Transparancy (FAccT). The title is taken from a conversation between Abeba Birhane, Chris Dancy and myself, where Chris offered the term State-of-the-Art White Supremacy.

the machine learning pipeline requires a deeper consideration of machine learning and the human embodiments within the pipeline. Moreover, it requires a deeper consideration of how machine learning, as an academic practice, presents itself and disembodies itself from the subjective human experiences that machine learning purports to be developed for. For this reason, I turn to considering how the machine learning pipeline embodies and disembodies human (experience). Therefore, I return to *RQ 1* by asking how subjective experiences are embodied in the machine learning pipeline (*RQ 4*) and what the implications of this are (*RQ 5*). In this chapter, I theorise on the core sources of these issues: the context within which models and exist, the models and the data. To describe these issues, I invoke the metaphor of the body in three different ways: first, pertaining to the physical material *human body* that we each possess; second, to signify a collection of observations and data points *created by humans*; and third, to refer subjective embodiments, that is how *social and cultural meaning* is embedded in the human experience and derivatives of it, i.e. data created by humans. I then reflexively apply my theory to the computational models in chapters 5 and 6 and provide a critique of these technologies through a consideration of the data generation process and the modelling stages of the machine learning pipeline. Finally, I discuss the implications of current practices in machine learning and argue that we must radically reconsider our current approaches to machine learning for social tasks, such that our approaches align better with the stated aims.

## 7.1 Disembodied Machine Learning

Machine learning is a practice that is concerned with making decisions based on machine-discernible patterns in observed data. Often, the data used to optimise machine learning methods are ‘extracted’ from the context within which they are created, i.e. by ‘scraping’ online platforms for user-generated content. Through this process of separating context and datum, a notion of ‘objectivity’ is imposed upon the data and the subsequent operations, namely, optimising machine learning methods on the data and their results further entrench this notion of objectivity. Datasets, or bodies of data, are thus created through the repetition of separating datum from context. These amalgamated bodies of data exist only by virtue of their strict separation from the material bodies from which the datum are derived. Such disembodied and amalgamated bodies are then used to optimise machine learning models. Machine learning methods come in two different forms: Supervised learning methods which seek to distinguish distinct limbs which are pre-drawn, e.g. classes, from the data; and unsupervised models which seek to identify discernible limbs of data within a single body of data without direct guidance from designers. For both supervised and unsupervised models the underlying data and the models applied to them strongly influence what bodies are discovered and what may be uncovered within these data. As Benjamin (2019) writes, technology operates within social structure “codes [that] operate within powerful systems of meaning that render some things visible, others invisible, and creates a vast array of distortions and dangers”.

With the advent of machine learning, a new technology came to be hailed as objective and unimpeded by subjective human biases, and by extension social marginalisation (O’Neil, 2017). However, an increasing amount of research suggests that social biases are common to machine learning models (Agarwal et al., 2018; Buolamwini and Gebru, 2018; Shah et al., 2020). Moreover, research has found that biases in the underlying data may be exacerbated by the machine learning models (Jia et al., 2020; Zhao et al., 2017). As a result of this growing awareness of the emergence of social biases in machine learning models, there has been a number of research directions seek to identify (Bender and Friedman, 2018; Buolamwini and Gebru, 2018; Mitchell et al., 2019; Shah et al., 2020), reduce or remove social biases (Agarwal et al., 2018; Jia et al., 2020; Romanov et al., 2019; Zhao et al., 2017) from machine learning models to prevent further marginalisation. However, such work assumes that social biases operate within a positivist logic which casts the removal of social biases as an optimisation problem. That is, this work assumes that bias is a finite and quantifiable entity that can be disentangled, isolated, and mathematically reduced out of the body of data or mathematical model, from which the designer, too, is disembodied.

Here, I provide a challenge to such a positivist logic. Drawing on work from feminist Science and Technology Studies and examples from Natural Language Processing, I argue that bias and subjectivity in machine learning pipelines are inescapable and can therefore not be simply be reduced or removed. Therefore, I hold that an ongoing recognition and reflection on our own positions, and the fiction of objectivity found in subjective realities is necessary to identify how the political choices are reflected throughout the machine learning pipeline. Through a conceptualisation of bias in these terms, I aim to shift the surrounding discourse away from bias an its elimination, to understanding subjective positionality and its implications on the machine learning pipeline from data generation to optimised model.

### 7.1.1 Embodiments in the Machine Learning Pipeline

Through Donna Haraway’s (1988) critique of objectivity (see chapter 2) it is possible to rethink how subjectivity is embedded in machine learning. Rethinking subjectivities in machine learning affords a recognition of machine learning’s potential to create social marginalisation without casting the problem in a positivist, optimisational logic. That is, we can come to understand the logics that govern machine learning for social data without casting the issue of discriminatory models as one of ‘debiasing’—a problem that purports to be an optimisable problem. In fact, by framing the issue of social bias away from such positivist fantasies, we are afforded the ability to view machine learning systems as technologies that are embedded in the very systems of oppressions that the models entrench. When we view machine learning systems as co-constitutive of the social systems within which they are embedded, it becomes clear that mathematical approaches to ‘debias’ such optimisation technologies reinforce a fantasy that issues of social discrimination and marginalisation are problems that can be treated as merely issues of statistics and mathematics, rather than living and lived histories of oppression. Thus, as the machine learning pipeline relies on data created by humans living within discriminatory social systems, the fantasy of ‘debiasing’ serves only to obscure how machine learning systems are complicit and co-constitutive of exclusion and marginalisation. I argue that the disembodied, or ‘objective’ position is expressed within the machine learning pipeline at multiple junctions:

1. In the data which is often removed from context and potentially adjudicated by externalised others,
2. in the model optimised on the disembodied data stemming from embodied data subjects, and
3. in the person designing the experiment and pipeline.

When constructing a dataset for machine learning, one must make a series decisions about how the data is to be constructed at different levels of granularity—from selecting a source of data to specific means of operationalising it. These decisions come to determine the ways in which contemporary machine learning methods disembody speakers from their speech. At a higher level, designers of machine learning infrastructures make decisions that impact every aspect of the pipeline. In their decisions, designers specify what counts and what does not count as relevant information, and how such information should be represented by machine learning models. Finally, once data has been gathered models are optimised on disembodied data from embodied subjects. In this way, the model becomes embodied through an amalgamation of limbs that have been disembodied from embodied data subjects. For instance, when constructing datasets for machine learning, including datasets for content moderation, it is necessary to make decisions on that delineate individual pieces of datum as relevant or irrelevant to the task across several layers of granularity. First, one must consider how to obtain a large sample of content which may contain the phenomena under study. In developing a resource for online abuse a decision must be made to which online communities, topics, or types of discourse may provide a large enough sample for study. The collected data is often produced by a large number of people on online platforms. Often, this process does not include collecting all posts produced by the individuals in the sample. Instead only posts that pertain to the phenomena under study are collected. In this way, a first step is made towards disembodying the sampled data from the individuals who have created it. In NLP, the primary focus of interest is the text, for which reason data about the user such as the name they provide (username and provided name), their location, and other meta data are often discarded. Thus, a second step is made towards disembodying the creator of the content, the speaker, from the speech that they produce. Moreover, the discursive structure, such as a posts and replies is often flattened, which further disembodies the speech act from the context within which it is produced. By making such decisions, data comes to be disembodied from the social and political contexts within which they are created.

Often an initial data sample which is large to ensure breadth in the sample is collected to obtain as much evidence towards the phenomena under study as possible. A second level of granularity in the data sample is then performed by selecting a smaller sample to study, within the larger sample. Here designers may seek to qualify and disqualify certain sub-samples in their originally collected sample, as some parts of the sample may not be pertinent or may only infrequently contain the phenomena under study, as the phenomena has conceptualised by the designers.

In the case of supervised machine learning, the data is passed through a third level of granularity. Here, the datum is provided to a number of annotators, who are rarely the creators of data in the sample. Moreover, it is the exception, rather than the rule, that the annotators are situated within the contexts of the creators of the data. These annotators then determine which limb of data i.e. the class, within the larger body of data, a given datum belongs to given a set of criteria for making such an adjudication. These criteria are developed and provided by the designers of the pipeline. Thus, the designers entrench their own subjectivities into the annotation process and exert control over it, and the annotators.

Turning to the optimisation technologies. Through ways in which they operation on data, machine learning models have different ways of embodying and disembodying data. In the optimisation process, machine learning models operate on disembodied data and further disembody them from the speakers through mathematical processes with the goal of settling on a distinct embodiment derived from the data. The specific way a machine learning model disembody data varies as a function of the specific mathematical functions that the models rely on, and seek to optimise. The disembodiment that the optimisation process performs happens through a manipulation of the data representations to draw discernible boundaries between the limbs of data, i.e. the classes. The underlying assumption that machine learning models make is that the data provided is all that is necessary to know to draw *meaningful* decision boundaries. It is then up to the designer to discern whether the decision boundaries drawn are truly meaningful or they represent spurious correlations. Making this decision however has proven to be large a challenge as recent research on the challenges of benchmarking highlights, that is evaluating the performance of machine learning models has proven to be a significantly challenging task due to a disconnect between what is measured in benchmark datasets and what the stated goals of a task, and benchmark, is (Bowman and Dahl, 2021; Kiela et al., 2021).

Finally, a great deal of attention has been given to how the lack of inclusion of designers across axes of identity can contribute to the producing socially biased systems (Holstein et al., 2019; West et al., 2019). However, the ways in which designers embed themselves in the machine learning pipeline can be opaque. I argue that designers come to embed their own subjectivities into the machine learning pipeline through choices that designers make in the process of developing these technologies, e.g., how to represent data, how features are selected and limits are set on vocabulary. In spite of being deeply embedded in the machine learning pipeline and technologies, designers are rarely subject to the machine learning systems, the harms of such systems, or are a part of the data that they rely on to create models.

In each of these aspects lay a large number of value judgements on the perspectives of data that are deemed relevant. I observe here a peculiarity of the machine learning pipeline. When data is disembodied from its creator, the data becomes an archive or a body of knowledge upon which the machine learning model draws on. In drawing upon the archive, machine learning models implicitly transform all positions that exist outside of the model's internal body, i.e. the archive become disembodied from the model. This transformation from disembodied to embodied then can serve as an explanation for calls for 'more' and 'more diverse' data (Holstein et al., 2019). It is worth noting here that the model-embodiment is tacitly acknowledged in the research fields of domain adaptation (Daumé III, 2007) and transfer learning. These fields acknowledge that to the information held in machine learning models is primarily applicable to the domains that are present in the datasets the models are optimised on, and that even small perturbations in the input to the model may drastically degrade its performance (Daumé III, 2007; Szegedy et al., 2014). These acknowledgements of embodiment exist in a self-contradictory tension with the position of objectivity within which these transfer-learning and domain adaptation methods operate within.

### 7.1.2 Embodiment in Data

As Gitelman (2013) argues, datasets do not exist naturally but must be produced. Considering this production of data through Haraway (1988), datasets can be understood as a form of knowledge that is produced through disembodying embodied experiences. Subjectivity can thus stem from a number of sources including the source of the data (Gitelman and Jackson, 2013), the data sampling method (Shah et al., 2020), and the selection of annotators (Derczynski et al., 2016; Talat, 2016).

Grounding my discussion in Natural Language Processing, I show how subjectivity manifests itself in machine learning models through a number of meaning-making processes, modelling choices, and data idiosyncrasies. I seek here to highlight the subjective and embodied nature of of data and classifications and that by taking a position of objectivity, we cannot do justice to the needs and wants of individuals or communities.

#### 7.1.2.1 Natural Language Processing Tasks

A range of, if not all, Natural Language Processing tasks are highly sensitive to the subjective values encoded in data. While such issues have frequently been studied in the context of high-level tasks, such as machine translation and abusive language detection, less attention has been given to core natural language processing tasks. Notably, the primary object of study of biases in core natural language processing has been the Part of Speech tagging task (Blodgett et al., 2016; Jørgensen et al., 2016) for which reason I also investigate the task. Generally, I argue that the adjudication of content, be it for abusive language, part of speech tagging, or any of the many other tasks that the field of natural language processing addresses delegitimises the very tools that are built, for users of said technologies due their presumed objectivity, which is neither truly neutral nor objective.

**High-level tasks** High-level tasks that require semantic and pragmatic understanding, e.g. machine translation, dialogue systems, metaphor detection, sarcasm detection, and abusive language detection are all highly sensitive to subjective values encoded in the data. In machine translation, research has identified a range of issues including stylistic (Hovy et al., 2020) and gender biases (Vanmassenhove et al., 2018). Particularly issues that pertain to the reinforcement of sexist stereotypes have been the object of academic (Zhao et al., 2017) and public (Locklear, 2018) scrutiny. A classic example of stereotypical translation are the translations stereotyped occupations from a language that does not contain grammatical gender to a language that does, e.g. the translations of *doctor* from English (unmarked for gender) to the German Arzt (marked for masculine) and *nurse* from English (unmarked for gender) to the German Krankenschwester (marked for feminine). Here we see that the 'objective', yet stereotyped translations are embodied in a patriarchal context which delegates high prestige to men and low prestige to women. While the translations may be correct in individual cases, they are not always correct. Assigning a single gold label to a given translation in itself provides an issue, as an input text may have several distinct and correct translations. However, most optimisation processes and evaluation algorithms assume that there exists a single correct translation, and that is the one the model is provided for optimisation and evaluation. The issue however is not always the adjudicated data that the model relies on. For instance, researchers have noted that word embeddings which are created through computing word co-occurrences similarly hold such stereotypical associations (Bolukbasi et al., 2016).

The issue of highly subjective 'truths' and gold labels for data extends to several other tasks such as text simplification and abusive language detection. In text simplification, numerous datasets make the claim that some words, sentences, or texts are difficult to read while others are easy. These labels are typically provided by human annotators who may agree on some labels. This agreement may in turn aid in the ability of models optimised to generalise to other datasets. However, the process of externalising the labelling process and disembodies the data, and subsequent models, from the embodiments of the diverse set of users of simplification technology, and how text difficulty manifests for them (Bingel et al., 2018).

Further, as is apparent in abusive language detection, the outcome of an annotation process, where the positionality of the adjudicators deviates within the group of adjudicators, may be less consistent annotations (Talat, 2016), which harms both the model and the supposed users of it. Many other causes and effects of disembodiment have been considered in the task of identifying abusive language. For instance, Talat et al. (2018) argue that datasets for abusive language embody a white perspective on respectability, finding that almost all uses of the *n-word* are tagged in the positive class in the dataset released by Davidson et al. (2017) regardless whether its use is within the African-American community. The labelling of the *n-word* does not necessarily embody a white perspective on respectability as the word does have frequent pejorative uses (Croom, 2013), however disregarding the usage of the word within the black diaspora, as datasets and tools frequently do (Davidson et al., 2019), does constitute a white supremacist idea of control of marginalised bodies and languages, for which there is a rich history (Craft et al., 2020). Indeed Talat et al. (2018) find that a large subset of the documents that contain the *n-word* in Davidson et al. (2017) that are labelled as hate speech and offensive language are likely to be in-group uses. This issue however is not limited to the dataset presented by Davidson et al. (2017), in fact, all datasets examined by Davidson et al. (2019) result in consistent and disproportionate error rates for African American English speakers. Systems built on these datasets, or as I argue here, datasets that are constructed within a social order where the white cisgender male body is constructed as the ‘neutral’ or ‘objective’, will replicate such biases. Thus, the race towards state-of-the-art machine learning models for content moderation is also a race towards state-of-the-art white supremacy.

**Core Natural Language Processing tasks** Beyond these issues existing in high-level tasks which may require a certain level of cognitive abstraction, they also exist in lower level, core natural language processing tasks such as Part of Speech tagging and dependency parsing. While I restrict the examples here to part of speech tagging, I contend that precisely the same arguments apply to dependency parsing.

Considering part of speech tagging, I find multiple junctions at which theory and data influence the process of developing tag-sets. First, the tag-set is developed based on a subjective linguistic theory that licenses some tags while rejects other. This linguistic theory is typically informed by observations on specific types language in the data it is developed to describe. Second, in the choice of sources of data. If the observed language production is a forum dedicated to computer games, the linguistic theories that form the basis of the tag-set are likely to focus on informal, and perhaps adolescent communication patterns. If on the other hand, the source of data primarily consists of newswire, the linguistic theory is likely to specifically address language production in formal settings. Third, in the development of a dataset of part of speech tags, I see similar issues of adjudication as for the high level tasks.[2] Thus, the development of part of speech tag-sets, and datasets it is applied on is a practice in developing descriptors and data which are mired in the context of the language production they seek to describe.

An example of one such tag-set is the Penn Treebank tag-set (Marcus et al., 1993), the *de-facto* standard for describing English word classes in natural language processing. The Penn Treebank tag-set was developed on primarily financial newswire text and published works in the United States of America in 1961 (Francis et al., 1982). The tag-set was primarily motivated by economic factors, such as there being several word classes that were ambiguous or word classes which occurred with such low frequency that they might only describe a single word. The Penn Treebank tag-set was thus developed with formal Standard American English in mind and is thus better suited to describe language which conforms to the English

[2] Although this may be mitigated by using trained linguists to label the dataset.

the underlying theory the tag-set describes than other varieties, sociolects, or slang (Blodgett et al., 2016; Jørgensen et al., 2016). This issue becomes even more drastically apparent when a tag-set developed for English is forced upon some other language, which it is far removed from being able to describe.

### 7.1.3 Embodiment in Modelling

While datasets are an important source of how a model may be embodied, machine learning methods themselves encode which embodiments are highlighted and which are subjugated. I primarily focus on supervised machine learning in my exploration of how models exacerbate disembodiment, as unsupervised methods are more directly volatile to subjective choices of the researcher, e.g. how the data is represented and which parameters the model is subject to.

As I seek to distinguish distinct model behaviours, I offer that models act on a spectrum from *localized* to *global* behaviour. In this conceptualisation, localized behaviour refers to when a model seeks to ground the datum within the context it is derived from, often using knowledge external to the optimisation data, e.g. context-aware models (Devlin et al., 2019; Garcia et al., 2019). Conversely, global modal behaviour instead operates only within the realm of the optimisation data it is optimised on, i.e. models that compound multiple senses of a word with little or no regard to their local contexts. Although language production is situated within a wider socio-political context of society, I limit my use of 'context' to mean the entirety of the sentence provided to the model.

By virtue of the subjective nature of embodying a datum within its context, there is large variation in how locally acting models may be developed. One tactic to situate datum within its context is through the use of transfer learning which allows for knowledge produced outside of the optimisation data to alter what the model embodies. For instance, should a dataset embody the language production within multiple sociolects, through the use of pre-optimised language models (Devlin et al., 2019) or mixed-member language models (Blodgett et al., 2016) deeper information about the sociolects in question can be provided by using the context of the sentence to identify how to situate the representation of a document.[3] The Multi-task learning paradigm also offers an avenue for embodying data in their contexts through their ability to encode information about the creator of the datum (Benton et al., 2017; Garcia et al., 2019). Transfer learning can similarly be applied to direct the model to embody the context a datum is derived from. For instance, Romanov et al. (2019) encode

[3] Different language and dialectal varieties may not be equally distributed in optimisation data for contextual models (Dunn, 2020), not unlike the issue of which bodies are given privileged space plague such models (Tan and Celis, 2019).

demographic information of the datum's creator into the model in efforts to deter models from learning stereotyped representations of marginalised speakers and communities.

Globally acting models on the other hand do not afford such embodiment. Through their reduction of a features in a model to a single sense, they are inherently unable to take into account the embodiment of the author, even if they are provided signals for how to embody a document at optimisation and inference time, due to the fact that such models remake meaning according to the distribution of features present at optimisation time. Any step taken towards embodying datum in its original context moves globally acting models along the spectrum towards being locally acting models. An example of such a step are word embeddings. Through their representation of words by the words that co-occur with the word's neighbouring words, thus assuming a similarity between the word and other words. While they provide a slight shift towards locally acting models through the frequency-based nature of how closely associated a word is, they fail to take a meaningful step away from being globally acting models, as all instances of a token occurring in the dataset will be reduced to a singular representation that does not take the surrounding context, i.e. the sentence, into account. It is important to note here that while word embeddings, and in fact contextual word embeddings provide a step towards localising models, the techniques of developing such embeddings rely on processes of disembodying a large set of data from their creators and constructing an amalgamated body of data that can collective embodiments. This amalgamated data carries with it many small influences of the specific subjective embodiments of each data creator.

### 7.1.4 The Embodied Designer

Though often referred to as a 'researcher' or 'developer', I draw on Herbert Simon (2019) to construct my understanding of a *designer*. I direct attention not to the profession of the individual or team in the machine learning pipeline but instead to the their action.

> Everyone designs who devises courses of action aimed at changing existing situations into preferred ones.
>
> (Simon, 2019, p. 111)

Following Simon (2019), the *designer* can then be understood to be anyone in the machine learning pipeline. While this includes annotators in addition to developers and researchers, I focus on the last two n their role as the designers as as they direct annotators and can supersede the choices made by the annotators.

The designer is embedded in the machine learning pipeline by virtue of the choices that they make throughout the development process, from the initial conceptualisation of the task to the final optimised system. All decisions that are described in section 7.1.2 and section 7.1.3 are either directly or indirectly made by the designer, in efforts to shape the final optimised model such that it fits the subjective positions of the designer. Direct decisions such as the choice of model, how to pre-process the data and transform it are direct decisions made by the designer. Indirect decisions refer to instances where the designer relinquishes control over some part of the process, for instance in annotation. Annotations are indirect decisions as the annotation guidelines are developed by the designer, yet the act of annotation is often performed by others. The decision on how the guidelines are to be operationalised however is a matter that is predominately controlled by the annotators, as they internalise and operationalise the annotation guidelines according to their own lived experiences and subjectivities. Moreover, the designer can choose several ways in which to disregard the data labelled by the annotators, should subsets of the annotations disagree with the positions that the designer holds. In this way, although designers relinquish some control through the annotation process, they maintain, and often exert power over the result of the annotation process. Through control of these decision making processes, the designers exert power and embody their own subjectivities into the machine learning pipeline.

An oft proposed solution to the issues of socially biased machine learning systems is to diversify the teams of designers who are developing the technologies (Holstein et al., 2019). This line of work has a similar argument to mine: That the subjective designers project an embodiment of self onto the technologies that they develop through the data and modelling choices that they make. Drawing on Haraway (1988), this suggests that the God trick that machine learning methods employ is a reflection of the ways in which the subjectivities of the designers are embedded in the systems. Rather than calling for diversifying the identities of the group of designers behind a tool, I argue that it is only through the recognition of ones own subjective embodiments that the issue of socially biased machine learning can be addressed. That is, it is only by recognising ones own subjectivities and actively making choices to represent the subjectivities of those that the technologies will be applied to, that one can hope to develop machine learning technologies that do not produce socially biased outcomes when applied to the target user group.

## 7.2 Embodiment and Disembodiment in the Abusive Language Detection Pipeline

In the above sections, I have described generally how subjective embodiments are manipulated and inserted throughout the machine learning pipeline for the general case. In this section, I turn my attention to abuse detection technologies in an examination of the subjective embodiments for this particular application of machine learning.

As with many machine learning pipelines, abusive language detection pipelines can have different starting points depending on whether any data is being annotated, or previously annotated data is used. For the latter case, the considerations of feature and model selection are particularly relevant to the development, however designers of models should be aware of the influences of subjective embodiment in the annotation process and as the effects of the annotation process remain in the dataset. One such effect of the designer of the dataset is that the subjective embodiments of the designers (and annotators) permeate through every step of the pipeline, as I have argued in the previous sections. For this reason, I address how the subjective embodiments influence each step of the abusive language detection pipeline in the subsections below.

### 7.2.1 Annotation Guidelines

Perhaps the most clear case of subjective embodiments being inserted into pipeline is in the annotation guidelines. For the abusive language detection there is no consensus on how to operationalise abuse (Talat et al., 2017). This lack of consensus leads distinct groups of designers to create their own guidelines on the basis of distinct sources and understandings of abuse. The choice of which background source is used depends strongly on the researchers. For instance, Talat and Hovy (2016) rely on critical race theory and gender studies to inform their annotation guidelines. Conversely, Davidson et al. (2017) rely on social media platforms' community guidelines to define abuse, and Fišer et al. (2017) rely on Slovenian legal definitions of hate speech to inform their annotation guidelines. These distinct motivations in part are informed by the cultures within which the researchers exist. For instance, the designers behind Davidson et al. (2017) are situated in the United States of American and their annotation guidelines are thus contextualised by the highly permissive freedom of speech protections enshrined by the second amendment of the constitution of the United States of America. The aim of their work, distinguishing hate speech from otherwise offensive content, can then be understood to be motivated by the issue of incorrectly labelling non-hateful entries as hateful, which could be read as contrastive to the freedom of speech protections in the United States of America. On the other hand, Talat and Hovy (2016) seek to address the issue of discriminatory speech, motivated by the harassment of women on social media. Their understanding of hate speech is then motivated by ensuring protection of marginalised communities, in part due to their belonging to a marginalised community. Thus, while annotation guidelines are strongly argued and motivated, the local embodiments and contexts of the authors influence the guidelines that they create.

### 7.2.2 Sampling Data

Beyond distinctions in the annotation guidelines, the sampling of data similarly is influenced by the subjective embodiments of the designers, resulting in distinct datasets examining different geographic cultures (Talat et al., 2018). Distinct motivations influence the questions that are under investigation in the research of different groups (Talat et al., 2018). For instance, Fišer et al. (2017) detail a framework based in the Slovenian legal context, where the authors of the study reside, directing the hate studied to be directly addressing hate occurring in Slovenia. Similarly, Davidson et al. (2017) seek to examine in hate in the United States of America, further they also limit their data sampling to tweets posted from inside the United States of America. Finally, Talat and Hovy (2016) specifically seek to address abuse towards women and other minorities, notably religious minorities and therefore do not limit the selection of data to any particular geography. In this way, datasets reflect more than investigations into different aspects of abuse. The dataset also reflect the specific interests and values of the designers as they choose sampling strategies that align with such interests. It is worth noting here that the source of funding for the construction of the pipeline may also hold influence. For instance, grants from government agencies may specify that abuse must be considered within a national context or geographic territory.

### 7.2.3 Annotators

Another source of the subjective embodiments that are encoded into the data is the annotation process itself (Talat, 2016). As Talat (2016) show, distinct groups of people will internalise and operationalise annotation guidelines according to their pre-existing values, that is their own subjectivities. As such, the resulting annotations embed how different people and groups operationalise the annotation guidelines. The outcome of annotation processes is then a mixture of designers' and annotators' subjectivities, written into data. This has strong implications for abusive language detection datasets, as these are the basis of models that encode annotators' views on acceptability, rather than abusive language directly. Unless annotators are carefully selected and educated, the annotations derived from groups of annotators may be internally incompatible within a dataset. Training and selection of annotators further provide space for designers to shift the annotators' work towards their own subjective positions on abusive language. Thus, annotators are also subject to the embodiments and goals of the designers. More specifically, such influence is is wielded by the designers through directly influencing aspects of annotations, such as the annotator selection (Talat, 2016), training of annotators (Vidgen et al., 2020a), the guidance that is provided (Palmer et al., 2020), the selection of annotations to use (Hovy et al., 2013), and through indirect selection criteria such as payment-level for annotation (Sabou et al., 2014).

As Hovy et al. (2013) show, the reliability of annotations is important to the successes of any subsequent task, however the question of what constitutes a reliable annotation is one that reflects the designer's positions on 'correct' labelling for a given task. In terms of abusive language detection, 'high quality' annotations thus reflect how the designers envision the task of abuse detection and the embodiments that the designers operate within. Consider for instance a pool of annotators with diverse and divergent political positions tasked with annotating hate speech. If the designers' understanding of what constitutes hate speech does not align with a sub-group of annotators, those annotations can then be disregarded and classed as "annotation errors'. However, considering the positionality of the divergent sub-group, their annotations may be entirely consistent with how they operationalise hate speech and their own subjectivities. That is, rather than errors, these annotations are simply embodying subjective positions that do not conform to those of the designers. For instance, should a group of people who politically self-identify to be on the far-right form a sub-group of annotators, then their operationalisation of hate speech is likely to diverge in key areas from the remaining annotating population, while being consistent with their own operationalisation of hate speech. In such a case, the designers are likely to disregard their annotation to ensure that the resulting data aligns with their own aims and subjectivities.

This issue exists not only for subsets of the annotation pool, entire pools of annotators may also consistently label within the designers' expectations, yet in conflict to the annotation framework. In one such instance, Davidson et al. (2017) find that "[h]uman coders appear to consider racists or homophobic terms to be hateful but consider words that are sexist and derogatory towards women to be only offensive". Such divergences in labels towards groups is inconsistent with the annotation guidelines provided by Davidson et al. (2017). However, the authors highlight this as a strength of their annotation framework, arguing that their annotation process allowed for distinguishing between hateful and offensive content, even if such distinction runs counter to the guidelines provided. Where there are distinct sub-groups within the data, selecting which group to consider has bearing on the internal consistency of the dataset and subsequently on any patterns a model might embody (Talat, 2016). This leads (Talat, 2016) to conclude that the selection of annotators should follow processes that allow for identifying, if not recruiting, annotators that share backgrounds and align on socio-political issues. Discrepancies between annotator backgrounds and political stance can also be addressed through annotator training, as Vidgen et al. (2020a) show. In instituting annotator training and addressing discrepancies between annotators, the designers directly train the annotators to reconstruct the embodied positions on hate speech that the designers hold. Thus, the designers wield direct and indirect influence over annotators and annotations, and hold power to elect whether the constructed dataset follows the embodiments of the annotators or of the designers themselves.

Finally, Sabou et al. (2014) argue that designing the task and setting the payment level can indirectly influence which annotators put themselves forward to work on the task. To attract ‘good’ annotators, it necessary to set payment for each annotation, using incentives such as high payment per document labelled. In this way, annotators are incentivised to learn the subjective positions of the designers. For a great deal of work in abusive language detection, the task and data are further disembodied in the annotation selection process as the annotators are unlikely to appear in the dataset. By adding an additional layer of disembodiment through the adjudication process that operates on already disembodied data, the annotation process further disembodies the data, and subsequently the model, from the context within which the data are derived from. One study however diverges from this notion of universal understandings of what abuse constitutes (Arora et al., 2020). By asking the very journalists who are a target of abuse to perform annotation work, they ensure that the labels that are associated with each data point is embedded within the subjective positioning of each journalist. This then affords training models that reflect the subjective positions that each journalist, who received abuse, have on online abuse. The task, in this ways moves away from being an abstract construction, to addressing the concrete needs of individual journalists.

### 7.2.4 Feature Selection

Considering what information the machine learning models consider to be pertinent, that is the bodies of data that uncovered through optimization, I similarly find ample space for subjective positioning. I construct here the notion of feature selection to mean the construction of features based on theoretical insights, hypotheses about the phenomena and the sub-selection from complete vocabularies. Considered through the lens of abusive language detection, harmful patterns of marginalisation are apparent through a selection bias, as designers realise themselves in the features that they construct.

#### 7.2.4.1 Manually Constructed Features

A large body of work on hate speech detection has investigated the question of which human constructed features are useful to the task of automated detection (Chiril et al., 2019a; Fortuna and Nunes, 2018; Stanković et al., 2020; Talat, 2016). Similarly, in chapter 5 I explore whether rationalising over content using LIWC can have beneficial influences for machine learning approaches for abusive language detection. Clearly, there is an interest in providing scaffolding for computational models to identify and address hate speech detection.

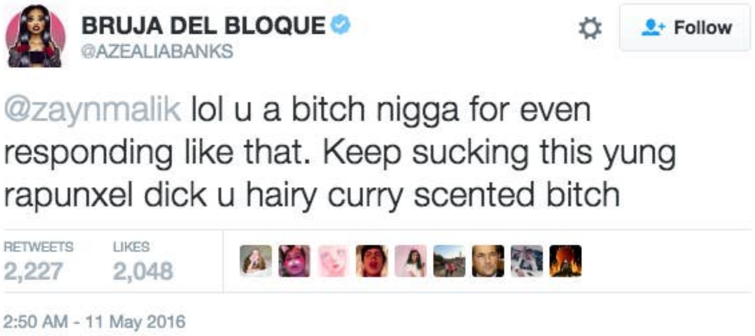
Through the use of higher level cognition, designers embed preconceived notions of what information computational models should deem relevant, for instance in chapter 5, I consider whether higher level cognitive information about the function of language can influence modelling and performance. The assumption is that while words may provide ample space for over-fitting models to specific instances and patterns that do not generalise beyond the data provided, other sources of information, i.e. the LIWC dictionary, may be less prone to over-fitting in such a way. By limiting the feature space to a much smaller discrete space of possible inputs, I argue that it is possible to achieve performance gains on out-of-domain data, relative to the input. Another frequently used modelling assumption is that computational models can benefit from considering words in some context, generally obtained using n-gram representations of the text (Chiril et al., 2019a; Davidson et al., 2017; Talat and Hovy, 2016). This modelling choice represents an assumption that that the context within which words appear carries significance beyond the word on its own. This stands in contrast to lexicon-based methods (e.g. Bassignana et al., 2019) that assume that the occurrence of some terms, disembodied from the local sentential context, should direct the model towards predicting either abuse or not abuse.

#### 7.2.4.2 Feature Selection in Neural Architectures

Many neural machine learning models are applied to text by providing the models with tokenised text, in which some minor replacements occur, e.g., substituting usernames, hashtags, and hyperlinks with stand-in tokens. This modelling choice made by the designers part relies on two strong assumptions. The first assumption is that all information seized from users will, to some degree be relevant to the modelling of abuse. The second assumption is that neural network models can use loss functions to update the model's internal representation of the data, in order to identify patterns that correlate in the input to the model with the output labels without the need for human cognition or oversight over the data or optimisation processes.

The first assumption stands in contrast with the use of externally computed word embeddings that neural network-based models frequently rely on (Isaksen and Gambäck, 2020; Kshirsagar et al., 2018). Such external word embeddings are most often computed at a much earlier time than the model is optimised, and their use requires that the vocabulary of the model is fixed to the vocabulary of the word embeddings. Thereby the use of pre-optimised word embeddings create a discursive shift between the knowledge contained in the data, and that which can be used by the models. That is, any shifts in language use and vocabulary that has occurred between the optimisation of the word embeddings and the optimisation of the model, will be either misrepresented or relegated to "unknown" tokens. In this alignment, such discursive shifts can have large impacts, particularly considering abusive language and hate speech on social media, where users frequently obscure their intended message (Röttger et al., 2020)—for example through intentional misspelling words. Such obscured tokens that are excluded from a model's knowledge and subsequent embodiment of data may in fact be key in distinguishing abusive content from the non-abusive.

Consider for instance the tweet posted by the American rapper Azealia Banks, an African American woman, directed towards fellow musician Zain Malik, a South Asian man, (see Figure 7.1). While the tweet uses profane language, the text is written in African American English, making the use of the *n-word* ambiguous. Similarly, as Azealia Banks is a woman, the use of the *b-word* similarly holds ambiguity, thus on the basis of those terms alone the tweet cannot unambiguously be identified as hate speech. While the tweet is clearly abusive and offensive, in part due to the call for Malik to perform a sexual act on Banks, it is only through the use of *curry scented* that the tweet moves unambiguously beyond *merely* being offensive to being hateful. As 'curry' and 'scented' are tokens likely to exist in pre-optimised word embeddings and language models, we might expect a model to correctly identify this tweet as abusive. However, as 'curry' and 'scented' are unlikely to frequently appear in context of abusive texts, the driver for a correct classification of hate speech is likely going to be the use of the *n-word* and the *b-word*—tokens that in this case cannot be relied on to determine abuse. Moreover, should there be attempts at obfuscating those tokens, e.g. by replacing all occurrences of the letter 'e' with the number '3' resulting in 'sc3nt3d', it is reasonably to expect that a language model and word embeddings would not have previously encountered this token. The token would then be transformed into an unknown token, and the hateful rhetoric would be unavailable to the model, forcing a model to rely on the ambiguous tokens to make a content moderation decision. On this basis, a model may incorrectly label it as simply offensive rather than hate speech, or correctly label it as hate speech, but for incorrect reasons.[4]



BRUJA DEL BLOQUE
@AZEALIABANKS

@zaynmalik lol u a bitch nigga for even responding like that. Keep sucking this yung rapunxel dick u hairy curry scented bitch

RETWEETS 2,227 LIKES 2,048

2:50 AM - 11 May 2016

**Fig. 7.1** Azealia Banks tweeting to Zain Malik.

The second assumption, that neural models can rely only on the input data and loss functions to identify relevant patterns without the need for human reasoning over the process or the relevance of the input data. This is lauded as a particular strength of neural network-based machine learning, as it is directed solely by the data without human interference. Through this data-driven process, models models construct and manipulate their own embodiment, on the basis of the disembodied data with which they are provided. Moreover, the designers' subjectivities are reflected in the construction and manipulation of the model's embodiment through the designers' decisions surrounding which data to include and how the model is constructed. In these decisions, the designers also make decisions on which, if any context is necessary to adequately represent the phenomena that is being modelled. Thus, designers construct and embed the normative values that determine relevance to a task, e.g., abusive language detection. That is, rather than theoretical or qualitative insights, model weights and probabilistic correlations are emphasised as the appropriate basis for classifications, so long as they reflect the designers' subjectivities. Such a practice therefore theorises that human cognition is rendered irrelevant by frequentist analyses of words and sub-words. This (implicit) assumption made by the designers contradicts recent studies that argue that language understanding models do not optimise to the point of having an ability to understand language, instead they optimise to parrot it (Bender and Koller, 2020).

Disregarding for a moment whether such models truly understand language or simply parrot it, what remains clear is that models that only use the surface forms of tokens lay on the globalised end of the model spectrum. The use of already-optimised language models

[4] Given the social biases against African American English in computational models, the tweet is likely to be identified as hateful in spite of the obscuring as a result of computational models disproportionately labelling African American English as hate speech (Davidson et al., 2019).

and word embeddings in a modelling architecture shift the models slightly towards a more localised end of the spectrum, as these allow for some social context to be derived from the way in text is written from a larger data sample. The use of these pre-optimised technologies thus come to shift the model embodiment away from the embodiments of the users and towards the embodiments of the designers' specific subjectivities. This shift happens as the choice of which language model and which word embeddings to use is a decision made by the designers. The decision is made on the basis of which specific pre-optimised technology best aligns with the designers' subjective position on what constitutes abuse and how it is best modelled, i.e. which underlying dataset for these technologies best aligns with the designers' perception of the distinctive features of abuse.

### 7.2.5 Model selection

As a number of models are optimised to identify which model best embodies the data, the designers must make normative decisions to identify what constitutes 'best'. In this process, designers make a final assertion, embedding their embodiments onto the decision on which model is selected for further use. However, the choice of designating what constitutes 'best' is often times a decision that is made prior to any model optimisation. For abusive language detection, best often refers to performance for some metrics. For instance, Gorrell et al. (2018) set out to have a model that has a high *precision* at the cost of *recall*. They make this choice to ensure high confidence in their model's predictions of the positive class as their use case is comments made to politicians, where the ability to criticise without sanction is of particular importance. Wulczyn et al. (2016) and Kshirsagar et al. (2018) on the other hand select their models using the Area Under the receiver operating characteristic Curve (AUC) and *F1-score*, respectively. Both of these metrics for measuring model performance give preference to models that balance classification error types, such that models are attuned to false positives as well as false negatives.

Through the choices of metrics, we can discern some aims of the modelling process. Where Gorrell et al. (2018) aim to situate their model within the context of abuse towards British Members of Parliament as it occurs on Twitter, they forego claims of universal applicability. The best performing model, within their understanding is a model which, within the context, produces as few false positives as possible, explicitly accepting that the number of false negatives may be high. Considering then the purpose of their modelling process, i.e. to allow for embodied downstream analysis of how abuse targets a very specific group, their choice affords an ability to speak to what is highly likely to be abuse within their understanding of abuse. On the other hand, their choice does not afford them the ability to speak to what is not abuse nor what their model misclassifies as not being abusive. Wulczyn et al. (2016) and Kshirsagar et al. (2018) on the other hand develop their models with the aim of obtaining a high degree of generalisation onto data outside of the sample that the model is optimised on. Within this goal lies an assumption that there exist a 'universal' and 'objective' understanding of what constitutes abuse, which is invariant to the specific embodiments of different people. That is, Wulczyn et al. (2016) and Kshirsagar et al. (2018) assume the existence of a global understanding of what constitutes abuse for an imagined average user, that is disembodied from all facets of human life.

## 7.3 Dissertation Models

Here I consider the two model types that I have developed for this dissertation, described in in chapter 5 and chapter 6, respectively. I document the considerations and assumptions that each model type reveals, and its implications for the machine learning pipeline. Rather than go through the entire pipeline, I begin my analysis at the entry points in the thesis, i.e., the choices of datasets and the modelling choices, as I exclusively use previously published datasets.

### 7.3.1 Vocabulary Reduction

In chapter 5, I optimised the machine learning models using the datasets published by Davidson et al. (2017), Talat (2016), Talat and Hovy (2016), Wulczyn et al. (2016), and de Gibert et al. (2018). The decision to use these datasets as optimisation data stems from these datasets originating from three distinct sources: Twitter; StormFront, Wikipedia editor discussions; and a white nationalist internet forum, respectively. To be able to measure the generalisability of the models optimised on Davidson et al. (2017), Talat (2016), and Talat and Hovy (2016), I reduce the multi-class classification tasks to binary classification tasks. Through this reduction in classes, I enforce a normative choice that the detection of abuse has greater value than the identification of the specific type of abuse, e.g., sexism or racism. My own experiences of hate speech and racialised abuse are at the heart of such a prioritisation, that is having been subject to such abuse I am more concerned with the ability to detect abuse than identifying which specific type of abuse it is. Further, the modelling choice of how to represent data are also subject to my subjectivities. While on one hand the reduction of the input space to a much smaller input space means that the size of the subsequent models, and by extension the complexity of the models, is greatly reduced.[5] On the other hand, through such a reduction in the vocabulary, a large majority of words will no longer be represented by the text in the models. Here, my belief that abuse detection models rely to strongly on token occurrences, ultimately impeding the goal of developing models that can protect marginalised people from abuse, is at the centre of my decision. On the other end of the vocabulary size spectrum, I use byte-pair encoded documents. Due to the nature of generating sub-words, this increases the size of the vocabulary in comparison to simply using the existing word. I use sub-words and byte-pair encoding to minimise issues of out-of-vocabulary items which may occur due to intentional obfuscation of words, e.g. through inserting spaces or punctuation in the middle of words (Röttger et al., 2020). This is also motivated by my lived experiences and observations of abuse towards others, where peple intentionally obfuscate words to circumvent the simple content filtering techniques, e.g., writing 'moslems' instead of 'Muslims'. While the modality I work with is text, such obfuscations also occur in the spoken word through intentional mispronunciation.

In the use of linear models as baseline models, the underlying assumption that I make is is that simple correlations of word occurrences with labels, are insufficient to capture the complex interactions between words that are required to make qualified judgements of abuse. This assumption too is influenced by my own positionality as a brown Muslim who grew up in a predominately white country where brown people, and in particular Muslims, are vilified for their existence. One such example were the police bulletins in Danish news while I was growing up. In these, the description 'Muslim looking' were routinely used to describe Brown men. Such experiences have made it clear to me that social norms surrounding the use of tokens cannot be readily understood from the words using simple correlations without greater contextualisation. For this reason, I use LSTMs as they can capture long interactions between words and are less directly reliant on the occurrence of individual patterns. I also use CNNs as a number of past studies having shown the efficacy of CNNs for abuse detection (Gambäck and Sikdar, 2017; Kolhatkar et al., 2020; Mitchell et al., 2019; Park and Fung, 2017; Rizwan et al., 2020; Safaya et al., 2020).

The models described in chapter 5 that only use words or byte-pair encoded words as input rely entirely on the optimisation data to optimise for patterns in data. Therefore, those models fall towards the very extreme of the globalised end of the model spectrum. The models that rely on the LIWC representations, although still on the globalised end of the spectrum, are further towards the localised end, as the LIWC dictionary is informed by considering data that is external to the optimisation data.

[5] I appreciate that even with a reduction of the model size and complexity, neural networks are still too complex to be readily understood without the aid of additional tools.

The motivation for using Byte-Pair encoding is reflected in a) my own personal assumptions about the importance of textual representations and b) computational considerations. I use BPE to minimise hte number of out-of-vocabulary items, as BPE deconstructs words in the optimisation data into smaller sub-words. This deconstruction affords minimising the influence of intentional obfuscations. Furthermore, the choice to BPE can aid the models in handling unknown tokens better.

Although some of the models are positioned further towards the localised end of the spectrum than others, all the models used in chapter 5 are on the globalised end of the model spectrum. Their globalised position derives from the fact that none of the data representations take local subjective positionalities of individuals in the data into account. Instead, all of the models rely on some abstraction away from the self through processes of disembodiment.

### 7.3.2 Multi-Task Learning

In chapter 6 I turn to the question of which constructs, in terms of machine learning tasks, would be helpful for machine learning models to embody to improve performance for a given abusive language detection task. Specifically, I examine whether jointly learning representations of sarcasm (Oraby et al., 2016), whether an argument is based in fact or feelings (Oraby et al., 2015), the moral sentiments elicited in tweets (Hoover et al., 2019), and related notions of hate speech and offensive language (Davidson et al., 2017; Talat, 2016; Talat and Hovy, 2016; Wulczyn et al., 2016) improves classification performance.

For this task, I reuse the BPE representations of documents from chapter 5. Therefore some of the embodiments of the models, with regard to text representation remain the same as described in section 7.3.1. Here I focus on the factors from chapter 6 that are distinct from chapter 5. As I include more datasets into consideration, I also implicitly invite the question 'why these datasets'? To answer this question, it's necessary to revisit the aims of each dataset.

One frequently identified issue with computational modelling of abuse is the issue of sarcasm (Röttger et al., 2020) and I use the dataset labelled for sarcasm that was proposed by Oraby et al. (2016). In this choice lay two assumptions: First, that computational models for abuse detection can benefit from better understanding what constitutes sarcasm. Second, that there does exist some overlap between sarcasm and abuse, where what appears to be abuse is in fact sarcastic. Both assumptions are the result of years of researching online abuse, and in particular exposing myself to the abuse that occurs in online spaces. While I may have become desensitised to abuse through the disproportionate amounts I am exposed to through my research, I frequently see that online abuse, and responses to it are expressed through humour, in particular sarcasm.[6]

The second dataset, asks the question of whether an argument is made in the basis of feeling or on the basis of facts (Oraby et al., 2015). As a majority of people who perpetrate online abuse do so infrequently (Talat and Hovy, 2016), an underlying cause for being abusive may be being impassioned, and thus being able to determine whether an argument is made with a basis in feelings or fact may be possible to help improve performance for abuse detection.

I also use a dataset annotated for moral foundation (Hoover et al., 2019). In this dataset, each document is labelled for which moral foundations it invokes in the annotators. Moral foundations and online abuse can be thought of as orthogonal concepts. Moral foundations, as annotated in the dataset, provide for a higher level cognition about the content that is read, in which abusive content is likely to elicit the moral sentiments that comprise the moral foundations framework (Hoover et al., 2019). I therefore believe that machine learning models jointly embodying abuse and moral foundations can aid with improving performance of machine learning classifiers for abuse detection. My own experiences of racialised abuse and observations of abuse in online spaces combined with the apparent desire of abusers to inflict harm upon their target are central to my inclusion of this task.

Finally, I use a number of datasets for online abuse (Davidson et al., 2017; de Gibert et al., 2018; Talat, 2016; Talat and Hovy, 2016; Wulczyn et al., 2016). For this task, I do not reduce the question of detecting to a binary task, instead I use the auxiliary task datasets as a means to provide the model with more conceptualisations of abuse. However, my subjective positioning does not change from what is detailed in subsection 7.3.1.

Similarly for the choices in developing the data and textual representations, my own subjective embodiments and experiences are a key factor in the modelling decisions (see section 7.3.1). This is particularly true for MTL, where I specifically set the weights for how much each task is to contribute to the main task through the frequency of selection. Such a weighting relies on my own consideration of how important each task is to the overall task of identifying abuse and, subsequently, the degree to which each auxiliary task should be afforded space to influence the model representations for abuse. The specific architecture of the model is influenced by its usefulness in prior work (Bingel et al., 2018) in addition to seeking an to answer the question of how more complex models would influence the performance on the task.

[6] By responses I mean general reactions and responses to abuse beyond the direct responses to a perpetrator of abuse.

Using the MTL framework has strong implications for where on the localisation spectrum the model is positioned. For instance, the use multiple different datasets to influence a single model precludes the extreme ends of the modelling spectrum. Jointly optimising multiple tasks in a shared internal model layer explicitly shifts models away from the extreme of the globalised spectrum, by providing using datasets that are external to the main task dataset and by contextualising the main task with the representations obtained from the auxiliary tasks. Similarly, the localised extreme of the model spectrum is also precluded using this method, as the auxiliary tasks do not afford embedding the subjectivities and embodiments of individuals. On a demographic level, however, MTL does hold potential for shifting towards the localised extrema, if and only if all auxiliary tasks also come from the demographic that the main task is concerned with. Moreover, as each auxiliary task will work, if as nothing else, as a regulariser for the main task, MTL will shift the model away away from the extremes of the spectrum. In my use of MTL for abusive language detection, with the auxiliary tasks that I have chosen, the models that I have developed are shifted away from the globalised extrema towards a more localised position on the model spectrum. However, as I do not optimise my models on any tasks that seek to make predictions on users, and the distinct datasets do not originate from the same demographic, the model remains a globalised model. Instead the models that I produce, by virtue of learning considerations on tone, argument basis, sarcasm, and moral sentiment, optimise for some representations of the faculties that I believe of importance to the task. These auxiliary tasks then provide an avenue for the models to be more closely situated within the how each individual person can be represented as a function of how they express themselves. Thus, while further towards being a localised end of the spectrum than the LIWC models, the models fall short of significantly situating the modelling of individuals within the context and lived experiences of that individual.

## 7.4 Discussion

Given that subjective choices and biases masquerading as disembodied ‘objective’ positions permeate the machine learning pipeline, the quest for objectivity and bias-free machine learning becomes redundant. This redundancy is made apparent as all choices in the machine learning development pipeline embody the subjective experiences of all who are a part of the pipeline, from the people whose data is seized, to annotators and the designers of the pipelines. As these experiences are embedded in the system, so slips away the illusion of ‘objectivity’ and ‘neutrality’ of the machine learning technologies. In fact, the search for objectivity in the pipeline creates a veneer of social progress that may cause further harm to already marginalised communities by obscuring and entrenching the dominance of certain bodies over others. Such harm is instituted by providing a veneer of more just, or fair, machine learning technologies that nonetheless perform institutional violence upon all who are externalised by the development process and the subjective experiences that lie at the heart of them. Without taking the unique embodiments of all data subjects into account, this imaginary of fair only serves as a justification of maintaining oppressive structures that are inherently harmful and reductive.

Considering the task of hate speech detection, developing automated tools, that are applied to a general population, makes inherent decisions on behalf of the user-group. The decisions made by the third party adjudicator, i.e. content moderation technologies, embed subjective experiences of the data subjects whose data has been stripped from the context it was created in, the annotators, and the designers of the technologies. Such decisions are codified through the machine learning pipeline, and are presented as disembodied and objective decisions on what constitutes hate speech. In this way, machine learning technologies embed normative socio-political positions on respectability and acceptability. These normative values come to be presented as ‘objective’ through the disembodiments that occur in the machine learning pipeline. However, such notions of objectivity merely provide a thin veil over the subjective embodiments of the designers and annotators. With the vast majority of research on abusive language detection being developed for English in the global north (Vidgen and Derczynski, 2020), the notions of respectability that are embedded in the technologies are normative for white majoritarian countries and cultures. Through such codification of white perspectives on respectability masqueraded as objective, attempts to address ‘bias’ in machine learning technologies for content moderation only serve to justify existing oppressive structures by further obscuring the subjectivities and norms embedded in the systems.

A consideration of how data is embodied can empower designers of machine learning systems by allowing them to reflect on what is embodied and how it is mired in context. Such considerations allow designers to interrogate the contexts within which data are created, and how meaning is made at each step in the dataset creation process. It is through such recognition of context and embodiment that one can realise that as contexts change, so does the applicability data. Further, only by such recognition of the deeply complex nature of embodiment and data can one hope to ask and ascertain which views the models privilege and which are subjugated. For building content moderation systems for the detection of hate speech and abuse, the designers of machine learning pipelines can ask how their own embodiments prejudice them to selectively sanction some speech patterns. Moreover, designers may want to ask themselves how such sanctions create downstream implications for the speech that is sanctioned.

Although there are methods with which we can move towards more localised machine learning models, what positions are given space remains a political question. It is only through wholly representing the context and embodiments of the data creator and the datum that one can hope to arrive at sufficiently localised models. Thus, rather than asking how to eliminate bias and subjective experiences from machine learning in the pursuit of objectivity, shifting the question to consider embodiments would ask us to reflect on the subjective experiences that are given voice. For hate speech detection, such reflections would have designers ask which groups understandings of abuse it is most appropriate to ground their definitions, and subsequently annotations and models, in. Such a shift would then require us to ask and reflect upon which bodies' subjective experiences we need to account for, such that we give voice to socially, and computationally marginalised groups.

Only by recognising the positionality of the designers and annotators of machine learning models and data, can one account for what (and whom) ones own position, and the models derived from it privilege and sanction, give space for and the political ramifications of this. For these reasons, it is imperative that machine learning moves away from consolidating power in the designers and move towards development practices that are rooted in the participation of the people who will be subject to the models, i.e. the intended users of the models. Participatory design practices however can quickly turn predatory if the turn to participatory design principles does not also provide for a redistribution of power (Kelly, 2019). Here, Sasha Costanza-Chock's (2018) work on design justice can provide a guide towards developing participatory design practices for machine learning. Costanza-Chock (2018) argues for design practices that centre the experiences and needs of the communities for whom the design practice is taking place. Specifically, Costanza-Chock (2018) argues that design practices should have "sustainable, community-led and -controlled outcomes." By working towards such goals, machine learning research can come develop processes and technologies that specifically address the needs of the communities for whom we are developing our technologies.

## 7.5 Summary

In this chapter, I have sought to examine how subjective embodiments permeate the machine learning pipeline, in efforts examine the machine learning infrastructures that underpin contemporary content moderation technologies. Thus, in this chapter I seek to examine *RQ I* by examining how subjective embodiments are embedded into machine learning infrastructures and what the consequences of such embodiments are. In this chapter, I then provide a reading of machine learning against the grain by critically examining how subjective embodiments become embedded within the machine learning infrastructure. By performing this reading, I have used this chapter to examine the ways in which machine learning systems come to produce socially biased outcomes, such that machine learning research can move beyond the discriminatory practices that we have developed.

Identifying processes for developing machine learning technologies that are not discriminatory, is of particular importance to content moderation technologies, as these technologies currently produce discriminatory outcomes in terms of censorship of marginalised communities (see chapter 4 for further details). The work that I have performed in this chapter, identifies specific ways in which machine learning, and machine learning for content moderation encodes the subjectivities that are widely read as social biases. To address this concern, I propose that researches be aware of the specific ways in which they embed their subjectivities throughout the machine learning pipeline and consciously make decisions in the development process that ensure that the subjectivities that are embedded within the systems reflect the aims and the subjectivities of the people that the content moderation systems are to be applied to. Specifically, I suggest that researchers are mindful of their own subjectivities and the desired outcomes of the technologies, and that research engages in a genuine efforts for participatory design by collaborating with the communities that technologies are developed for. Moreover, I argue that the universalist notions applied in machine learning, including the algorithmic detection of abusive language, contribute strongly to the ongoing marginalisation that machine learning systems perpetuate. Finally, I argue that it is only by taking steps away from such universalist notions and towards co-developing systems with communities that are community-led and community-controlled that we can hope to overcome the issues of discriminatory systems and, in the case of content moderation, systems that do not censor marginalised communities.

# Chapter 8

# Conclusion

In this thesis I have sought to explore the content moderation infrastructures that are built for classifying textual abuse in online spaces. The contributions of the thesis are structured around four central themes: How the notions of 'healthy' and 'toxic' content are operationalised, the implications of such operationalisation and how these come to embody hegemonic imaginaries on respectability; how large vocabulary reductions, that represent the mental and emotional states of speakers rather than their words influence the ability of models to classify in-domain and out-of-domain data; how different, apparently related, tasks can be used to jointly optimise model representations to gain models that more closely come to reflect the contexts a given speaker is operating in when speaking; and finally, how the subjective embodiments of data subjects and modellers alike are embodied in the machine learning pipeline and how these collectively come to privilege hegemonic discourses. These four distinct themes are connected through two over-arching research questions:

**RQ i** *What technical and social factors are present in the socially discriminatory predictions of content moderation systems?*

**RQ ii** *In which ways can computational methods be used to address limitations that are influential in discriminatory outputs from computational modelling?*

To adequately answer these questions, I address each theme in turn through through multiple disciplines. This approach affords insights into the technical, social, and political dimensions of content moderation infrastructures for abusive texts. The contributions in this thesis are thus in part theoretical and in part experimental in nature. By examining the questions through theoretical and experimental lenses, I can begin to uncover the political and technical complexities of content moderation infrastructures Furthermore, this multi-disciplinary approach affords insights that are opaque when the questions are addressed purely theoretically or purely experimentally approach. The thesis has been structured such that I start and finish with primarily theoretical contributions while the primarily experimental contributions constitute the middle of the thesis. I choose this structure to remain faithful to the machine learning pipeline for content moderation, addressing first definitional questions and then questions of modelling. Finally, I take a step back and reflect on the machine learning pipeline from start to an end.

In efforts to answer *RQ I* and *RQ II*, I formulate sub-questions that ask the following directed research questions:

**RQ 1** *How are notions of 'toxicity' operationalised and modelled, and what are their socio-political implications for content moderation systems?*
**RQ 2** *What are the modelling implications of using LIWC to substitute the use of words and sub-words as input tokens?*
**RQ 3** *How do the individual and combinatory use of abuse classification and non-abusive tasks impact classification of specific forms of abuse?*
**RQ 4** *How are the subjective embodiments embedded in the machine learning pipelines?*
**RQ 5** *What are the implications of such subjective embodiments with regard to developing machine learning models?*

In chapter 4, I address research question 1. In addressing *RQ 1*, I critique of how notions of 'toxicity' are operationalised and employed through a consideration of of Mary Douglas (2005) work on social pollution. I argue that notions of 'toxic' and 'healthy' are operationalised within several socio-cultural contexts: the context of the designers of the task, the contexts of the annotators, and the contexts of distinct modelling techniques. These contexts, however are not reflected on in the process of developing automated content moderation tools, instead they are assumed to have little influence, resulting in models that collapse each of these contexts into a single entity that embodies them. Such embodiment is predicated on efforts towards obtaining a global understanding of 'toxic' and 'healthy' content, thus reproducing racialised and gendered positions on respectability. As such, the content moderation systems that I examine, engage in toxic slippage, where they simultaneously over-police the content produced by people inhabiting marginalised positions while also failing to protect these groups from 'toxic' content.

Moreover, I argue that the model development process exhibits three prominent avenues that lead to models that maintain pre-existing, hegemonic power structures. First, through the optimisation data, which is likely to more frequently have occurrences of some identity terms in the positive class than in the negative class, e.g. 'Black', 'gay', and 'woman'. Thus, a model is likely to embody that identity terms have a greater association to 'toxic' content. Further, I argue that even when several datasets with competing definitions of abuse are used, the optimisation procedures result in models better representing the overlaps of data and labels, at the expense of where the datasets or labels diverge from one another. Second, models come to embody discriminatory norms through the use of pre-optimised embeddings that take a distributional perspective on language production and are known to harbour harmful social biases against minoritised groups (Speer, 2017). Lastly, I argue that the models themselves are likely to exacerbate hegemonic positions, given that machine learning models have been shown to amplify social biases that exist in datasets (Zhao et al., 2017). These findings have an impact on future work for computational modellers as they provide theoretical scaffolding for why and how marginalised communities come to suffer under abuse classification models. Thus, the contributions here begin to forge a path towards abuse detection models that are centred around the experiences of those who are most likely to suffer harms from misclassifications.

One challenge that is raised in chapter 4 is the notion of the pluralist model that Opt Out develop. While a pluralist model does allow for people embed their own subjectivities into the modelling process, they are not exempt from the critiques made in chapter 4. In fact, as we see in the chapter, Opt Out incorrectly moderates fig. 4.4 due to the uses of the *b-word* and references to sexual promiscuity. Moreover, as pluralist models also need some centralisation for optimising the machine learning model, they are similarly subject to the risks of optimising towards hegemonic positions. This risk may be slightly decreased as the pluralist model is only be applicable to a single person at a time. Thus the hegemonic positions that they may come to embody might only affect singular individuals who specifically provide data to the model that would encode such hegemonic positions. This further raises an issue for consideration, namely that of the harms individuals may enact on a larger community. Specifically, if a person's model comes to embody positions that are tailored to the individual, are there any limitations that should be set for minimal notions of acceptability, that are applied for everyone using such pluralist models? For instance, should people who believe in the genocide of other people be afforded the ability to determine that calls for genocide is acceptable content while resistance to such calls is deemed as unacceptable? If such minimal notions of acceptability are to be set, further questions around who is to determine them and what exceptions should be made to this remain as vital questions.

In chapter 5, I turn towards research questions 2. Starting with *RQ 2*, I optimise three different neural network models and two baselines for five different datasets, each experimenting with 3 different data representations: Word tokens, sub-words, and LIWC encoded documents. I find that using LIWC encoded documents to optimise machine learning models can obtain competitive results with linear models and neural networks that are optimised on word tokens and sub-word representations alike. Specifically, I find that for linear models, LIWC-based models can obtain classification performances that are as good or better than models optimised on the other two representations, though for some dataset and model combinations there is a significant drop in performance. For neural network models, I similarly find that LIWC models can provide for in-domain classification improvements in some cases, and in most cases provide with a competitive performance. I find that all three neural model types are well suited for the use of LIWC, with the best performances achieved by the CNN models. Thus, I conclude that while there is space for the improvements in model performances, LIWC based modelling can provide for an alternative to using word-token or sub-word document representations.

I further observe how machine learning models are affected by the vocabulary change in terms of the time required to fully optimise them. I find that in many cases, LIWC-based modelling for neural networks show a reduction in the time it takes for a model to finish the optimisation procedure. However, I also observer that in some cases, the LIWC-based models take as long, or longer, than models optimised on other document representations. In particular, I find a relationship between the complexity of the model and the model optimisation time, where the more complex a model is, the longer it takes to optimise. That is, I find that all but one LIWC MLPs take less time to complete the optimisation procedure than their word-token and sub-word counterparts. For CNNs, LIWC models tend to finish optimising close to as close as word-token models, and faster than sub-word based models, with a single exception where the LIWC model takes almost twice as long as its closest competitor. Finally, for LSTM models, I similarly find that in most configurations of datasets, LIWC-based models tend to finish optimising quicker than the other models. Here too there is a single outlier in which the LIWC LSTM model takes longer to finish optimising than all other models. Thus, I find that in terms of speed in optimising the models, LIWC-based modelling in most cases provides for as fast or faster model optimisation procedures.

I also examine how the LIWC-based models perform when evaluated on out-of-domain data by applying on all models across all datasets, including those that the model has not been optimised for. Using this method, I find that LIWC-based models often provide for out-of-domain performances that out-perform all other out-of-domain models. Particularly

LIWC-based models optimised on the *Toxicity* dataset performs well on out-of-domain data. Finally, I observe that all models tend to perform better on out-of-domain datasets where the goal of the dataset resembles the goal of the dataset the model is optimised on.

As the results of my experiments show, there is space for improving in-domain and out-of-domain classification performances by thinking carefully about how data is represented. Research question2 invites researchers to think carefully about data representations. Moreover, in specific next steps, there is space to investigate what the impact on modelling is when LIWC-based document representations are combined with and word-token or sub-word token representations. Such combinations would allow for using models that have pre-optimised embedding layers for the word and sub-word parts of the data representations. The use of pre-optimised layers is likely to provide for additional improvements on both in-domain and out-of-domain performances.

Turning questions of context, I address research questions 3 in chapter 6. In this chapter, I experiment with Multi-Task Learning with three main tasks for distinct forms of abuse, the *Offence* detection task, the *Hate Speech* detection task, and the *Toxicity* detection task. Each of these is also used as an auxiliary task when it is not the primary task. For auxiliary tasks I use the three main task datasets in turn, and the *Hate Expert* dataset, as my abusive auxiliary tasks. For my non-abusive auxiliary tasks I use the *Sarcasm* detection task, the *Moral Sentiment* prediction task, and the *Argument Basis* task. Through my experiments, I find that there is a positive impact on model performances, in terms of improvements over a single task MLP baseline when using almost any dataset as and auxiliary task. Specifically, I observe that the three main task models that I optimise all benefit from using sarcasm detection as an auxiliary task. Moreover, two out of three auxiliary tasks also benefit from using the remaining two non-abusive auxiliary tasks. Using combinations of only non-abusive tasks also improves modelling performances, though none of these combinations achieve the best-performing auxiliary task combination.

In fact, I find that using abusive tasks as auxiliary tasks has a positive impact on the model performances. Unlike the non-abusive tasks there is one auxiliary task setting for one dataset where the highest performance is achieved by a combination of only abusive auxiliary tasks. The use of abusive tasks in particular has a beneficial impact on performances in terms of *precision* score. In the abusive tasks however, I also find that not all tasks are equally suited. In particular, I observe a relationship between abusive datasets that share similarities in either data source or in annotation goals. For instance, the *Hate Expert* dataset only performs well enough as an auxiliary tasks to be selected as for further when the main task is the *Hate Speech* task which was sampled from the same collection of data and annotated using the same guidelines. Moreover, I find that the *Offence* auxiliary task is useful for both main tasks, where it shares the dataset source with the *Hate Speech* task and shares annotation goal with the *Toxicity* task.

When considering combinations of abusive and non-abusive auxiliary tasks, I find that combining abusive and non-abusive auxiliary tasks provides for some of the best model improvements. Specifically, I find that using the *Sarcasm* task in conjunction with one or more abusive tasks provides for high performing models. This suggests that encoding representations for sarcasm detection in combination with an auxiliary abuse dataset can have benefits on the main task performance. It is worth noting that all three non-abusive auxiliary tasks appear in the best-performing auxiliary task configurations and the best performances are obtained when abusive and non-abusive auxiliary tasks are used together.

The findings in this chapter, provides several paths for future work. For instance, one avenue for future work is to explore more non-abusive auxiliary tasks such as sentiment analysis. Additionally, future research can address improvements in modelling e.g. by using more complex modelling architectures or by optimising the weight each auxiliary task is given.

Finally, in chapter 7, I address the final two research questions: *RQ 4 & 5*. In this chapter, I read against the grain in my considerations of the machine learning pipeline for NLP. I further apply my insights to models developed in this thesis for abuse detection.

Through my reading, I find that subjectivity is embedded into machine learning in all processes that humans with their subjective experiences are involved in. The subjective experiences of people involved in the modelling pipeline express their subjective experiences through the data collection, annotation, and model building processes. Thus, I argue that to address issues of bias, fairness, and representing the users of machine learning models, it is necessary for awareness of the subjective experiences that modellers want to represent and develop processes which foster the human and computational expression of these. One avenue for such development processes is through participatory design with a focus on notions of design justice, as outlined by Costanza-Chock (2018). Moreover, to develop which are fair an equitable, it is necessary to start with the development process of machine learning models with those whose experiences are not embodied in machine learning. Thus, rather than attempting to force models that have been optimised to reproduce oppressive structures, this research calls for modelling to be centred around the subjective, lived experiences of people and their needs.

In future work, there is more space for practitioners and researchers to engage in identifying how their own subjective experiences influence their design decisions. Moreover, an implication of this work is that models that embody desired subjective experiences can be developed, given that the resources for this are developed.

Returning to the guiding questions of how machine learning systems for content moderation come to produce socially discriminatory outputs and the ways in which computational methods can come to address some of these concerns, my findings in this thesis are that machine learning systems very poorly represent the subjective experiences of large groups of people, and the computational approaches that I developed to more faithfully represent these people make positive steps but still fall shy of making truly faithful representations. Moreover, content moderation technologies for online abuse expressed through text at present have largely not sought to closely represent the subjective experiences of users. Particularly, they have failed to represent the perspectives and experiences of those who stand to be harmed most by content moderation technologies, instead focusing on goals such as having models that take a global perspective on abuse. These issues are the result of a computing culture that seeks to abstract away subjectivity in search of, if not 'objective truths', global consensus on inherently subjective questions. However, as I show in this thesis, there is vast, and largely unexplored, space for developing models that more closely seeks to represent the people and thus better make space for peoples subjective experiences. For instance, in chapter 5, we see how using modelling that seeks to encode the mental and emotional state of the author yields for improved performances on out-of-domain data, when the annotation goals of the in-domain and out-of-domain data align, thus providing space for generalising specific perspectives. Moreover, as observed in chapter 6 the use of MTL can allow for models to optimise representations of related auxiliary tasks, such as whether a comment is sarcastic or the expressed moral sentiments, can allow for deeper engagements with the intentions of the speaker, thus moving modelling to more closely represent the speakers and their intentions.

In seeking answers to the guiding questions, my work in this thesis both fails to achieve the research goals and manages to fulfil them. My work fails to achieve these goals by not fully engaging with the questions in more depth and in the same vein achieves these research goals by providing a step into these questions. Moreover, my work achieves the research goals by providing alternative readings and methods for machine learning that can afford more faithful representations of people. On the other hand, the work fails to achieve its goals by not providing definitive answers, but instead provides suggestions for research directions. While I construct the successes and failures in absolute terms here, my position is more nuanced. The work that I have performed provides for some beginnings of research directions and for some further steps for pre-existing directions. The failures of not achieving definitive answers are offset by having identified new questions to ask while the successes of providing steps for new directions also make space for analyses of the limitations of the directions that I have investigated.

In this thesis, I have sought to identify challenges and opportunities for developing models for the content moderation of online abuse. This is scaffolding for a hopeful path forward - one that centres the subjective experiences and humanity of those impacted by content moderation technologies, and that brings back considerations of the social hierarchies that make them most vulnerable to abuse. I hope to bring the experiences of these people back into the heart of the task.

## 8.1 Limitations and Future Work

The work in this thesis has a number of limitations, both computationally and theoretically. A core limitation that applies to all aspects of this thesis is that the work is primarily investigating content moderation as it applies to the global economic north, specifically for English speaking nations, with a particular focus on the United States of America. For all aspects of this thesis, future work would be well suited to develop on the work presented here by centring specific contexts in the global economic south, where content moderation, or the lack thereof has had disastrous consequences. Moreover, my chapter on content moderation and social pollution (see chapter 4) provides a next step, expanding on Liboiron et al. (2018). While this chapter extends a general theoretical framework, it is centred around how content moderations deals with race and gender, and to a lesser degree sexuality in the global economic north. A limitation here is then that the considerations and how they apply to other contexts, i.e. the global economic south are not included. Future work could then think more deeply about the specificities with which our framework requires extension for specific contexts and that of content moderation in the global economic south.

In terms of computational limitations, the work in chapter 5 is limited to English as the LIWC dictionary is only defined for English. Moreover, using LIWC is most appropriate for content written in mainstream American English, as this is the only variant of English that it is defined for. Finally, the use of psychometrics has deep-rooted issues, a reliance on any psychometrics carries a risk of reproducing harmful and hegemonic reductions of psychological constructs. Future work could then consider other forms of low-vocabulary representations that are not rooted in such problematic histories. Furthermore, future work could consider how low-vocabulary representations can be used in conjunction with pre-optimised technologies such as large language models and word embeddings. Centring other languages than English would also have the direct benefit of requiring a different vocabulary reduction technique as LIWC only exists for mainstream American English.

Similarly to the limitations of chapter 5, the work I have performed on Multi-Task Learning would benefit greatly from centring other languages than English and other populations and nations beyond that of the United States of America. Moreover, although there are improvements on the baselines in this chapter, there is ample space for considering how MTL could be used to improve results further. This space also includes considerations of wider computational architectures, e.g. the inclusion of pre-optimised technologies as internal layers of the model, or directly experiment with MTL using solely large language models. Although I perform an extensive investigation of tasks, several more tasks could be considered for exploration including rumour detection and sentiment analysis.

Both computational chapters could be extended along the lines of directly including considerations of demographic belonging. That is, the loss functions for the neural networks used in both chapters could take into account the demographic information that is available about the speakers. By considering demographic information, machine learning models could come to start to encode pre-existing power structures and account for these in their functions. The final contribution of my thesis is largely theoretical and functionally translation work between fields. For this reason, introducing the notion of disembodiment to the machine learning and NLP communities, starts at the beginning, and thus the chapter seeks to provide a foundation for future thought. The limitations of this work then is that there is far more depth to consider in how each of the aspects highlighted contribute to disembodying machine learning technologies from human experiences. Moreover, the pipeline that I have sought to read and analyse is a research pipeline. In a production pipeline for a commercial entity, machine learning technologies are often embedded in deeper technical structures. As machine learning technologies are increasingly being deployed into such production pipelines, it is prudent with a consideration of how embodiment and disembodiment happens in a commercial, for-profit entity.

Finally, with the work in this thesis I have sought to lay the foundations for new directions for abusive language detection and machine learning. The scaffolding that I provide directly invites and welcomes work to build around the scaffolding and develop the structures further, such that we can emphasise justice in the processes we create for developing new technologies for people and communities.

# References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://www.aclweb.org/anthology/L16-1704`.

Ibrahim Abu Farha and Walid Magdy. Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL `https://www.aclweb.org/anthology/2020.osact-1.14`.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/agarwal18a.html`.

Daniela Agostinho, Catherine D'Ignazio, Annie Ring, Nanna Bonde Thylstrup, and Kristin Veel. Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive. *Surveillance & Society*, 17(3/4):422–441, September 2019. ISSN 1477-7487. doi: 10.24908/ss.v17i3/4.12330. URL `https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/12330`.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76, 2018. doi: 10.1109/ASONAM.2018.8508247.

Julia Angwin and Hannes Grassegger. Facebook's secret censorship rules protect white men from hate speech but not black children, 2017. URL `https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms`. Accessed: 16/06/2019.

Naomi Appelman, Joanna M.L. van Duin, Ronan Fahy, Joris van Hoboken, Natali Helberger, and Brahim Zarouali. *Access to Digital Justice: In Search of an Effective Remedy for*

*Removing Unlawful Online Content*, chapter 14. Edward Elgar, 2021. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3961390`. To be published as Chapter 14 in X. Kramer et al., Frontiers of Civil Justice (Edward Elgar, forthcoming 2022).

David S. Ardia. Free speech savior or shield for scoundrels: An empirical study of intermediary immunity under section 230 of the communications decency act. *Loyola of Los Angeles Law Review*, 43(2):373–506, 2010.

Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 7–15, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.2. URL `https://www.aclweb.org/anthology/2020.alw-1.2`.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4914-7. doi: 10.1145/3041021.3054223. URL `https://doi.org/10.1145/3041021.3054223`.

James Banks. Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3), 2010.

Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, page 1–6. CEUR-WS, 2018.

Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018)*, 2019.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2126. URL `http://aclweb.org/anthology/S17-2126`.

Lori G. Beaman. The myth of pluralism, diversity, and vigor: The constitutional privilege of protestantism in the united states and canada. *Journal for the Scientific Study of Religion*, 42(3):311–325, 2003. doi: 10.1111/1468-5906.00183. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-5906.00183`.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL `https://www.aclweb.org/anthology/Q18-1041`.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://www.aclweb.org/anthology/2020.acl-main.463`.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL `https://doi.org/10.1145/3442188.3445922`.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Ruha Benjamin. *Race after technology: abolitionist tools for the new Jim code*. Polity, Medford, MA, 2019. ISBN 9781509526437.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-1015`.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL `https://www.wandb.com/`. Software available from wandb.com.

Joachim Bingel and Johannes Bjerva. Cross-lingual complex word identification with multitask learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 166–174, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-0518. URL `https://www.aclweb.org/anthology/W18-0518`.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, 2018.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1120. URL `https://www.aclweb.org/anthology/D16-1120`.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL `https://www.aclweb.org/anthology/2020.acl-main.485`.

Violet Blue. Suicide, violence, and going underground: Fosta's body count, 2018. URL `https://www.engadget.com/2018/04/27/suicide-violence-and-going-underground-fosta-sesta/`. Accessed: 15/06/2019.

Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, and Yaser Al-Onaizan. Neural Word Decomposition Models for Abusive Language Detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3515. URL `https://www.aclweb.org/anthology/W19-3515`.

Robert J. Boeckmann and Jeffrey Liew. Hate speech: Asian american students' justice judgments and psychological responses. *Journal of Social Issues*, 58(2):363–381, Jan 2002. ISSN 0022-4537, 1540-4560. doi: 10.1111/1540-4560.00265.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016.

Léon Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Yves Lechevallier and Gilbert Saporta, editors, *Proceedings of COMPSTAT'2010*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD. ISBN 978-3-7908-2604-3.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4843–4855. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.385. URL `https://www.aclweb.org/anthology/2021.naacl-main.385`.

Andre Brock. Deeper data: a response to boyd and Crawford. *Media, Culture & Society*, 37(7):1084–1088, October 2015. ISSN 0163-4437, 1460-3675. doi: 10.1177/0163443715594105. URL `http://journals.sagepub.com/doi/10.1177/0163443715594105`.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, February 2018. PMLR. URL `http://proceedings.mlr.press/v81/buolamwini18a.html`.

Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

Judith Butler. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, 1990.

Elinor Carmi. *Media distortions: understanding the power behind spam, noise, and other deviant media*. Digital formations. Peter Lang, 2020. ISBN 978-1-4331-6691-4.

Rich Caruana. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, ICML'93, pages 41–48, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. ISBN 1-55860-307-7. event-place: Amherst, MA, USA.

Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997. ISSN 08856125. doi: 10.1023/A:1007379606734. URL `http://link.springer.com/10.1023/A:1007379606734`.

Yunfei Chen, Lanbo Zhang, Aaron Michelony, and Yi Zhang. 4is of social bully filtering: Identity, inference, influence, and intervention. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2677–2679, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398723. URL `http://doi.acm.org/10.1145/2396761.2398723`.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. Multilingual and Multitarget Hate Speech Detection in Tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 351–360, Toulouse, France, 2019a. ATALA. URL `https://www.aclweb.org/anthology/2019.jeptalnrecital-court.21`.

Patricia Chiril, Farah Benamara Zitoune, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. Multilingual and multitarget hate speech detection in tweets. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II: Articles courts*, page 351–360. ATALA, Jul 2019b. URL `https://aclanthology.org/2019.jeptalnrecital-court.21`.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1271. URL `https://www.aclweb.org/anthology/P19-1271`.

Jennifer Cobbe. Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*, October 2020. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-020-00429-0. URL `http://link.springer.com/10.1007/s13347-020-00429-0`.

Communications Decency Act of 1996. 47 U.S.C. § 230, 1996.

Sasha Costanza-Chock. Design justice, a.i., and escape from the matrix of domination. *Journal of Design and Science*, Jul 2018. doi: 10.21428/96c8d426. URL `https://jods.mitpress.mit.edu/pub/costanza-chock/release/4`.

Justin T. Craft, Kelly E. Wright, Rachel Elizabeth Weissler, and Robin M. Queen. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, 6(1):389–407, 2020. doi:

10.1146/annurev-linguistics-011718-011659. URL https://doi.org/10.1146/annurev-linguistics-011718-011659.

Kate Crawford and Tarleton Gillespie. What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428, 2016.

Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist eory and antiracist politics. *University of Chicago Legal Forum*, 1989(1), 1989.

Adam M. Croom. How to do things with slurs: Studies in the way of derogatory words. *Language & Communication*, 33(3):177 – 204, 2013. ISSN 0271-5309. doi: https://doi.org/10.1016/j.langcom.2013.03.008. URL http://www.sciencedirect.com/science/article/pii/S0271530913000232.

Alessandro Acquisti Daegon Cho. The more social cues, the less trolling? an empirical study of online commenting behavior. In *Proceedings of the Twel h Workshop on the Economics of Information Security (WEIS 2013)*, 2013.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P07-1033.

Thomas Davidson, Dana Warmsley, Micheel Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*, 2017.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3504. URL https://www.aclweb.org/anthology/W19-3504.

Simone de Beauvoir. *The Second Sex*. Knopf, New York, 1953.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5102. URL http://aclweb.org/anthology/W18-5102.

Varty Defterderian. Fair housing council v. roomates.com: A new path for section 230 immunity. *Berkeley Technology Law Journal*, 24(1):563–592, 2009.

Leon Derczynski, Sean Chester, and Kenneth Bøgh. Tune your brown clustering, please. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, page 110–117. INCOMA Ltd. Shoumen, BULGARIA, Sep 2015. URL https://aclanthology.org/R15-1016.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1111.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 25(2):700–732, April 2021. ISSN 1095-5143, 1936-4822. doi: 10.1007/s12119-020-09790-w. URL http://link.springer.com/10.1007/s12119-020-09790-w.

Ángel Díaz and Laura Hecht. *Double Standards in Social Media Content Moderation*. 2021. URL https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. *The Social Mobile Web*, 11(02), 2011.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.23. URL https://www.aclweb.org/anthology/2020.emnlp-main.23.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA, December 2018. ACM. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278729. URL https://dl.acm.org/doi/10.1145/3278721.3278729.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. Multi-task learning using AraBert for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 97–101, Marseille, France, May 2020. European Language Resource Association. ISBN 979-10-95546-51-1. URL https://www.aclweb.org/anthology/2020.osact-1.16.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the*

*Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1166. URL `https://www.aclweb.org/anthology/P15-1166`.

Mary Douglas. *Purity and danger: an analysis of concept of pollution and taboo*. Routledge classics. Routledge, London ; New York, 2005. ISBN 978-0-415-28995-5. OCLC: ocm50333732.

Kevin Drakulich, Kevin H. Wozniak, John Hagan, and Devon Johnson. Race and policing in the 2016 presidential election: Black lives matter, the police, and dog whistle politics. *Criminology*, 58(2):370–402, May 2020. ISSN 0011-1384, 1745-9125. doi: 10.1111/1745-9125.12239.

W.E.B. Dubois. *Black Reconstruction in America: an essay toward a history of the part which black folk played in the attempt to reconstruct democracy in America, 1860-1880*. Philadelphia, Penn. : Albert Saifer, 1935.

Jonathan Dunn. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 2020. doi: 10.1007/s10579-020-09489-2.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low Resource Dependency Parsing: Cross-lingual Parameter Sharing in a Neural Network Parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2139. URL `http://aclweb.org/anthology/P15-2139`.

European Commission. Code of conduct on countering illegal hate speech online. Technical report, European Commission, 2016.

European Commission. Regulation of the european parliament and of the council on preventing the dissemination of terrorist content online. Technical report, European Commission, 2018.

European Union Agency for Fundamental Rights. Proposal for a regulation on preventing the dissemination of terrorist content online and its fundamental rights implications: Opinion of the european union agency for fundamental rights. Technical report, European Union Agency for Fundamental Rights, 2019.

Facebook. Community Standards Enforcement, n.d. URL `https://transparency.facebook.com/community-standards-enforcement#hate-speech`.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org, 2018. URL `http://ceur-ws.org/Vol-2150/overview-AMI.pdf`.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3007. URL `https://www.aclweb.org/anthology/W17-3007`.

Luchina Fisher and Brian McBride. "ghostbusters" star leslie jones quits twitter after online harassment, 2016. URL `https://abcnews.go.com/Entertainment/ghostbusters-star-leslie-jones-quits-twitter-online-harassment/story?id=40698459`.

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL `https://doi.org/10.1145/3232676`.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, May 2021. ISSN 03064573. doi: 10.1016/j.ipm.2021.102524. URL `https://linkinghub.elsevier.com/retrieve/pii/S0306457321000339`.

James Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An Intersectional Definition of Fairness. *arXiv:1807.08362 [cs, stat]*, September 2019. URL `http://arxiv.org/abs/1807.08362`. arXiv: 1807.08362.

Electronic Frontier Foundation. Section 230 of the communications decency act, n.d. URL `https://www.eff.org/issues/cda230`. No Date. Accessed: 15/06/2019.

Winthrop Nelson Francis, Henry Kucera, Henry Kučera, and Andrew W Mackie. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin, 1982.

Stuart Frost. 'A Bastion of Colonialism': Public Perceptions of the British Museum and its Relationship to Empire. *Third Text*, 33(4-5):487–499, September 2019. ISSN 0952-8822, 1475-5297. doi: 10.1080/09528822.2019.1653075. URL `https://www.tandfonline.com/doi/full/10.1080/09528822.2019.1653075`.

Vijaya Gadde and David Gasca. Measuring healthy conversation, 2018. URL `https://blog.twitter.com/en_us/topics/company/2018/measuring_healthy_conversation.html`.

Philip Gage. A New Algorithm for Data Compression. *C Users J.*, 12(2):23–38, February 1994. ISSN 0898-9788. Place: USA Publisher: R & D Publications, Inc.

Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics, 2017. URL `http://aclweb.org/anthology/W17-3013`.

Noa Garcia, Benjamin Renoust, and Yuta Nakashima. Context-aware embeddings for automatic art analysis. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 25–33, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367653. doi: 10.1145/3323873.3325028. URL `https://doi.org/10.1145/3323873.3325028`.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, 2018. URL `http://arxiv.org/abs/1803.09010`. arXiv: 1803.09010.

Katharine Gelber and Luke McNamara. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341, May 2016. ISSN 1350-4630, 1363-0296. doi: 10.1080/13504630.2015.1128810. URL `http://www.tandfonline.com/doi/full/10.1080/13504630.2015.1128810`.

Ysabel Gerrard. Social media content moderation: six opportunities for feminist intervention. *Feminist Media Studies*, 20(5):748–751, July 2020. ISSN 1468-0777, 1471-5902. doi: 10.1080/14680777.2020.1783807. URL `https://www.tandfonline.com/doi/full/10.1080/14680777.2020.1783807`.

Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich, and Sarah Myers West. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4), October 2020. ISSN 2197-6775. doi: 10.14763/2020.4.1512.

Debbie Ging and Eugenia Siapera. Special issue on online misogyny. *Feminist Media Studies*, 18(4):515–524, Jul 2018. ISSN 1468-0777, 1471-5902. doi: 10.1080/14680777.2018.1447345.

Lisa Gitelman, editor. *"Raw data" is an oxymoron*. Infrastructures series. The MIT Press, Cambridge, Massachusetts, 2013.

Lisa Gitelman and Virginia Jackson. Introduction. In Lisa Gitelman, editor, *"Raw Data" Is an Oxymoron*, pages 1–14. MIT Press, Cambridge, Massachusetts, 2013.

Yoav Goldberg. *Neural network methods for natural language processing*. Number 37 in Synthesis lectures on human language technologies. Morgan & Claypool Publishers, San Rafael, 2017. ISBN 978-1-62705-295-5 978-1-68173-235-0 978-1-62705-298-6. OCLC: 990794614.

Eric Goldman. The complicated story of fosta and section 230. *First Amendment Law Review*, 17(Symposium):279–293, 2018.

Hila Gonen and Yoav Goldberg. Lipstick on a Pig:. In *Proceedings of the 2019 Conference of the North*, pages 609–614, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL `http://aclweb.org/anthology/N19-1061`.

G. Gorrell, M. A. Greenwood, I. Roberts, D. Maynard, and K. Bontcheva. Twits, Twats and Twaddle: Trends in Online Abuse towards UK Politicians. In *Proceedings Of The Twelfth International Conference On Web And Social Media*, California, United State of America, June 2018. Association for the Advancement of Artificial Intelligence. URL `http://www.aaai.org/Library/ICWSM/icwsm18contents.php`.

Jessica Guynn. Facebook while black: Users call it getting 'zucked,' say talking about racism is censored as hate speech, 2019. URL https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/. Accessed: 16/06/2019.

Stuart Hall. The spectacle of the other. In *Representation: Cultural representations and signifying practices*, volume 7. Sage London, 1997a.

Stuart Hall. Race, the Floating Signifier, 1997b. URL https://shop.mediaed.org/race-the-floating-signifier-p173.aspx.

Donna Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575–599, 1988. ISSN 0046-3663. doi: 10.2307/3178066. URL https://www.jstor.org/stable/3178066.

Benjamin Heinzerling and Michael Strube. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Patricia Hill Collins. *Black Feminist Thought*. Routledge, 0 edition, Jun 2002. ISBN 978-1-135-96014-8. doi: 10.4324/9780203900055. URL https://www.taylorfrancis.com/books/9781135960148.

Gabriel Emile Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. A longitudinal measurement study of 4chan's politically incorrect forum and its effect on the web. In *ICWSM*, volume abs/1610.03452, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL https://direct.mit.edu/neco/article/9/8/1735-1780/6109.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300830. URL https://doi.org/10.1145/3290605.3300830.

Home Office. Action against hate the uk government's plan for tackling hate crime. Technical report, United Kingdom Home Office, 2016.

bell hooks. *Talking back: thinking feminist, thinking black*. Between the Lines, 1st ed edition, 1989. ISBN 978-0-921284-08-6.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. Moral Foundations Twitter Corpus: A Collection of 35k Tweets Annotated for Moral Sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, November 2020. ISSN 1948-5506, 1948-5514. doi: 10.1177/1948550619876629. URL http://journals.sagepub.com/doi/10.1177/1948550619876629.

Joseph Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida M Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment, Apr 2019. URL psyarxiv.com/w4f72.

Dirk Hovy and Shannon L. Spruit. The Social Impact of Natural Language Processsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL http://aclweb.org/anthology/P16-2096.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1132.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. “you sound just like your father” commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.154.

Theresa Ingram. Opt Out Tools, 2020. URL https://www.optoutools.com/.

Vebjørn Isaksen and Björn Gambäck. Using Transfer-based Language Models to Detect Hateful and Offensive Language Online. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 16–27, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.3. URL https://www.aclweb.org/anthology/2020.alw-1.3.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. Mitigating gender bias amplification in distribution by posterior regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.264. URL https://www.aclweb.org/anthology/2020.acl-main.264.

Jigsaw. Perspective API, 2017. URL https://github.com/conversationai/perspectiveapi/.

Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. Automated Identification of Verbally Abusive Behaviors in Online Discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3505. URL `https://www.aclweb.org/anthology/W19-3505`.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1130. URL `https://www.aclweb.org/anthology/N16-1130`.

Mladen Karan and Jan Šnajder. Cross-Domain Detection of Abusive Language Online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5117. URL `http://aclweb.org/anthology/W18-5117`.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, July 2018. URL `http://proceedings.mlr.press/v80/kearns18a.html`.

Janet Kelly. Towards ethical principles for participatory design practice. *CoDesign*, 15(4): 329–344, Oct 2019. ISSN 1571-0882, 1745-3755. doi: 10.1080/15710882.2018.1502324.

Erin M. Kerrison, Jennifer Cobbina, and Kimberly Bender. “Your Pants Won’t Save You”: Why Black Youth Challenge Race-Based Police Surveillance and the Demands of Black Respectability Politics. *Race and Justice*, 8(1):7–26, January 2018. ISSN 2153-3687, 2153-3687. doi: 10.1177/2153368717734291. URL `http://journals.sagepub.com/doi/10.1177/2153368717734291`.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4110–4124. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.324. URL `https://www.aclweb.org/anthology/2021.naacl-main.324`.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Jaime Lee Kirtz, Zeerak Talat, Christine Tomlinson, and Wendy Hui Kyong Chun. Definitely maybe: The Impact of the Specificity of Ambiguity in Content Moderation. *In Press*, 2022.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1288`.

Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. Classifying Constructive Comments. *First Monday*, In press(In press), August 2020. URL `http://arxiv.org/abs/2004.05476`.

April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. Detecting cyberbullying: Query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 195–204, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1889-1. doi: 10.1145/2464464.2464499. URL `http://doi.acm.org/10.1145/2464464.2464499`.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, page 34–43. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.alw-1.5. URL `https://www.aclweb.org/anthology/2020.alw-1.5`.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. Predictive Embeddings for Hate Speech Detection on Twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 26–32, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5104. URL `https://www.aclweb.org/anthology/W18-5104`.

Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 177–188, Barcelona Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372853. URL `https://dl.acm.org/doi/10.1145/3351095.3372853`.

Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 1621–1622. AAAI Press, 2013. URL `http://dl.acm.org/citation.cfm?id=2891460.2891697`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=H1eA7AEtvS`.

Mary Graw Leary. The indecency and injustice of section 230 of the communications decency act. *Harvard Journal of Law & Public Policy*, 41(2):553–622, 2018.

Josh Lepawsky. No insides on the outsides. *Discard Studies*, 2019. URL `https://discardstudies.com/2019/09/23/no-insides-on-the-outsides/`.

Max Liboiron, Manuel Tironi, and Nerea Calvillo. Toxic politics: Acting in a permanently polluted world. *Social Studies of Science*, 48(3):331–349, June 2018. ISSN 0306-3127, 1460-3659. doi: 10.1177/0306312718783087. URL `http://journals.sagepub.com/doi/10.1177/0306312718783087`.

Frederick Liu and Besim Avci. Incorporating Priors with Feature Attribution on Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1631. URL `https://www.aclweb.org/anthology/P19-1631`.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. Datasets of Slovene and Croatian Moderated News Comments. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5116. URL `http://aclweb.org/anthology/W18-5116`.

Emma Llansó, Joris van Hoboken, Paddy Leerssen, and Jason Harambam. Artificial intelligence, content moderation, and freedom of expression. *Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*, 2020. URL `https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf`.

Mallory Locklear. Google is working to remove gender bias in its translations, 2018. URL `https://www.engadget.com/2018-12-07-google-remove-gender-bias-translations.html`. Accessed on 11/07/2020.

Emilia L. Lombardi, Riki Anne Wilchins, Dana Priesing, and Diana Malouf. Gender violence. *Journal of Homosexuality*, 42(1):89–101, 2002. doi: 10.1300/J082v42n01_05.

Audre Lorde. *The Master's Tools Will Never Dismantle the Master's House*, page 110–113. Crossing Press, c2007, 1984. ISBN 978-1-58091-186-3.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv:1711.05101 [cs, math]*, January 2019. URL `http://arxiv.org/abs/1711.05101`. arXiv: 1711.05101.

Rijul Magu, Kshitij Joshi, and Jiebo Luo. Detecting the hate code on social media. *arXiv preprint arXiv:1703.05443*, 2017.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1993. URL `https://www.aclweb.org/anthology/J93-2004`.

Rob Marvin. How Google's Jigsaw Is Trying to Detoxify the Internet. *PC Magazine*, January 2019. URL `https://in.pcmag.com/gallery/128319/how-googles-jigsaw-is-trying-to-detoxify-the-internet`.

Adrienne Massanari. #gamergate and the fappening: How reddit's algorithm, governance, and culutre support toxic technocultures. *New Media & Society*, 2015.

Lance T. McCready. Understanding the Marginalization of Gay and Gender Non-Conforming Black Male Students. *Theory Into Practice*, 43(2):136–143, May 2004. ISSN 0040-5841, 1543-0421. doi: 10.1207/s15430421tip4302_7. URL http://www.tandfonline.com/doi/abs/10.1207/s15430421tip4302_7.

Peggy McIntosh. White privilege and male privilege: A personal account of coming to see correpondences through work in women's studies. 1988.

Johannes Skjeggestad Meyer and Björn Gambäck. A Platform Agnostic Dual-Strand Hate Speech Detector. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 146–156, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3516. URL https://www.aclweb.org/anthology/W19-3516.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the nternational Speech Communication Association (INTERSPEECH 010)*, volume 2010, pages 1045–1048. International Speech Communication Association, 2010. ISBN 978-1-61782-123-3. URL http://www.fit.vutbr.cz/research/view_pub.php?id=9362.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, Atlanta, GA, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Chandra Talpade Mohanty. Under western eyes: Feminist scholarship and colonial discourses. *boundary 2*, 12/13:333–358, 1984. ISSN 01903659, 15272141. URL http://www.jstor.org/stable/302821.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3512. URL https://www.aclweb.org/anthology/W19-3512.

Ella Myers. Beyond the Psychological Wage: Du Bois on White Dominion. *Political Theory*, 47(1):6–31, February 2019. ISSN 0090-5917. doi: 10.1177/0090591718791744. URL https://doi.org/10.1177/0090591718791744.

Karsten Müller and Carlo Schwarz. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, page jvaa045, October 2020. ISSN 1542-4766, 1542-4774. doi: 10.1093/jeea/jvaa045. URL https://academic.oup.com/jeea/advance-article/doi/10.1093/jeea/jvaa045/5917396.

Sarojini Nadar. "stories are data with soul" – lessons from black feminist epistemology. *Agenda*, 28(1):18–28, Jan 2014. ISSN 1013-0950, 2158-978X. doi: 10.1080/10130950.2014.871838.

Radford M. Neal. *Bayesian learning for neural networks*. Ph.D., University of Toronto, Toronto, Canada, 1996. URL http://www.cs.toronto.edu/~radford/ftp/thesis.pdf.

Victor Nina-Alcocer. AMI at IberEval2018 Automatic Misogyny Identification in Spanish and English Tweets. In Paolo Rosso, Julio Gonzalo, Raquel Martínez, Soto Montalvo, and Jorge Carrillo de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 274–279. CEUR-WS.org, 2018. URL http://ceur-ws.org/Vol-2150/AMI_paper8.pdf.

Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor. *Computational Linguistics*, 46(2):487–497, June 2020. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00379. URL https://direct.mit.edu/coli/article/46/2/487-497/93368.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883062. URL http://dx.doi.org/10.1145/2872427.2883062.

Safiya Umoja Noble. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, New York, 2018. ISBN 9781479849949 9781479837243.

Cathy O'Neil. *Weapons of math destruction: how big data increases inequality and threatens democracy*. Penguin Books, London, 2017. ISBN 978-0-14-198541-1. OCLC: 994642027.

Shereen Oraby, Lena Reed, Ryan Compton, Ellen Riloff, Marilyn Walker, and Steve Whittaker. And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 116–126, Denver, CO, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0515. URL https://www.aclweb.org/anthology/W15-0515.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles, September 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3604. URL https://www.aclweb.org/anthology/W16-3604.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4674–4683, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1474. URL https://www.aclweb.org/anthology/D19-1474.

Alexis Palmer, Christine Carr, Melissa Robinson, and Jordan Sanders. Cold: Annotation scheme and evaluation data set for complex offensive language in english. *Journal of Linguistics and Computational Linguistics, Special Issue*, to appear(to appear):tbd, 2020.

Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, page 363–370. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-2051. URL `https://www.aclweb.org/anthology/P19-2051`.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6): 102360, Nov 2020. ISSN 0306-4573. doi: 10.1016/j.ipm.2020.102360.

Luciana Parisi. *Techno ecologies of sensation*, chapter 10, pages 182–197. Palgrave Macmillan, 2009. ISBN 978-0-230-52744-7.

Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3006. URL `http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-3006.pdf`.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 2799–2804. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1302. URL `http://aclweb.org/anthology/D18-1302`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count. Technical report, Erlbaum Publishers, 2001. URL `https://www.researchgate.net/publication/246699633_Linguistic_inquiry_and_word_count_LIWC`.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC2015. Technical report, University of Texas at Austin, Austin, Texas, 2015.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

B. T. Polyak and A. B. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM J. Control Optim.*, 30(4):838–855, July 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046. Place: USA Publisher: Society for Industrial and Applied Mathematics.

Lisa Posch, Arnim Bleier, Fabian Flöck, and Markus Strohmaier. Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics. *arXiv:1812.05948 [cs]*, December 2018. URL http://arxiv.org/abs/1812.05948. arXiv: 1812.05948.

Lutz Prechelt. *Early Stopping - But When?*, volume 1524 of *Lecture Notes in Computer Science*, page 55–69. Springer Berlin Heidelberg, 1998. ISBN 978-3-540-65311-0. doi: 10.1007/3-540-49430-8_3. URL http://link.springer.com/10.1007/3-540-49430-8_3.

Siouxsie Q. Anti-sex-trafficking advocates say new law cripples efforts to save victims, 2018. URL teenvogue.com/story/fosta-sesta-anti-sex-trafficking-bill. Accessed: 15/06/2019.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1482. URL https://aclanthology.org/D19-1482.

Mohammed Rafi Arefin. Abjection: A definition for discard studies. *Discard Studies*, 0(0), September 2019. URL https://discardstudies.com/2019/09/23/no-insides-on-the-outsides/.

Jacquelyn Rahman. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171, 2012. doi: 10.1177/0075424211414807.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.394. URL https://www.aclweb.org/anthology/2020.acl-main.394.

Marek Rei. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2121–2130, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1194. URL https://www.aclweb.org/anthology/P17-1194.

K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244, Dec 2011. doi: 10.1109/ICMLA.2011.152.

Roopika Risam. Toxic femininity 4.0. *First Monday*, March 2015. ISSN 1396-0466. doi: 10.5210/fm.v20i4.5896. URL https://journals.uic.edu/ojs/index.php/fm/article/view/5896.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. Hate-Speech and Offensive Language Detection in Roman Urdu. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2512–2522, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.197. URL `https://www.aclweb.org/anthology/2020.emnlp-main.197`.

Sarah T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. ISBN 9780300235883.

Aja Romano. A new law intended to curb sex trafficking threatens the future of the internet as we know it, 2018. URL `https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom`. Accessed: 15/06/2019.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1424. URL `https://www.aclweb.org/anthology/N19-1424`.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Michael Beißwenger, Michael Wojatzki, and Torsten Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum, sep 2016.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. Hatecheck: Functional tests for hate speech detection models, 2020.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/497_Paper.pdf`.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online), December 2020. International Committee for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.semeval-1.271`.

Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio. Detecting nastiness in social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 63–72. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3010. URL `http://aclanthology.coli.uni-saarland.de/pdf/W/W17/W17-3010.pdf`.

Magnus Sahlgren, Tim Isbister, and Fredrik Olsson. Learning Representations for Detecting Abusive Language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 115–123, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5115. URL http://aclweb.org/anthology/W18-5115.

Haşim Sak, Andrew Senior, and Françoise Beaufays. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv:1402.1128 [cs, stat]*, February 2014. URL http://arxiv.org/abs/1402.1128. arXiv: 1402.1128.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 12th International AAAI Conference on Web and Social Media, ICWSM 2018, pages 330–339. AAAI press, January 2018.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://www.aclweb.org/anthology/P19-1163.

Guy Schaffer. Queering Waste Through Camp. *Discard Studies*, February 2015. URL https://discardstudies.com/2015/02/27/queering-waste-through-camp/.

E. F. Schumacher. *Small is beautiful: economics as if people mattered.* Harper Perennial, 1973. ISBN 978-0-06-091630-5.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL http://aclweb.org/anthology/P16-1162.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.468. URL https://www.aclweb.org/anthology/2020.acl-main.468.

Dara Sharif. We got 'zucked': Facebook censors speech that calls out racism, black activists charge, 2019. URL https://www.theroot.com/we-got-zucked-facebook-censors-speech-that-calls-out-1834286507. Accessed: 16/06/2019.

David R. Sheridan. Zeran v. aol and the effect of section 230 of the communications decency act upon liability for defamation on the internet. *Albany Law Review*, 61(1):147–180, 1997.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive Language and Hate Speech Detection for Danish. *arXiv:1908.04531 [cs]*, August 2019. URL http://arxiv.org/abs/1908.04531. arXiv: 1908.04531.

Caty Simon. On backpage, 2018. URL http://titsandsass.com/on-the-death-of-backpage/. Accessed: 15/06/2019.

Herbert A. Simon. *The sciences of the artificial*. The MIT Press, Cambridge, MA, third edition [2019 edition] edition, 2019. ISBN 978-0-262-53753-7.

Vinay Singh, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay, and Manish Shrivastava. Aggression Detection on Social Media Text Using Deep Neural Networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 43–50, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5106. URL http://aclweb.org/anthology/W18-5106.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Sara Sood, Judd Antin, and Elizabeth Churchill. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1481–1490. ACM, 2012a.

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. Using crowdsourcing to improve profanity detection. In *AAAI Spring Symposium: Wisdom of the Crowd*, volume SS-12-06 of *AAAI Technical Report*. AAAI, 2012b.

Robyn Speer. How to make a racist AI without really trying, July 2017. URL http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/.

Ranka Stanković, Jelena Mitrović, Danka Jokić, and Cvetana Krstev. Multi-word Expressions for Abusive Speech Detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, online, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.mwe-1.10.

Luke Stark. Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48 (2):204–231, Apr 2018. ISSN 0306-3127, 1460-3659. doi: 10.1177/0306312718772094.

Megan Stevenson. Megan Stevenson on Twitter: "I just realized something about the AUC as a measure of predictive accuracy. It's usually dependent on the heterogeneity of the underlying sample! Like, if you wanted to predict who will earn 6 figure salaries based on parental income, your predictions will be much more... 1/2" / Twitter, 2021. URL https://twitter.com/MeganTStevenson/status/1382071900751429634.

Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. Detecting harassment in real-time as conversations develop. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 19–24, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3503. URL `https://www.aclweb.org/anthology/W19-3503`.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL `https://www.aclweb.org/anthology/P19-1355`.

Kitty Stryker. What the fosta/sesta anti-sex trafficking bill means, 2018. URL `https://www.teenvogue.com/story/fosta-sesta-anti-sex-trafficking-bill`. Accessed: 15/06/2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the Importance of Initialization and Momentum in Deep Learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1139–III–1147. JMLR.org, 2013. URL `http://proceedings.mlr.press/v28/sutskever13.pdf`. event-place: Atlanta, GA, USA.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA, 2014. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2969033.2969173`.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. Studying Generalisability across Abusive Language Detection Datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1088. URL `https://www.aclweb.org/anthology/K19-1088`.

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5): 44–54, May 2013. ISSN 0001-0782, 1557-7317. doi: 10.1145/2447976.2447990. URL `https://dl.acm.org/doi/10.1145/2447976.2447990`.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, January 2014. 2nd International Conference on Learning Representations, ICLR 2014 ; Conference date: 14-04-2014 Through 16-04-2014.

Zeerak Talat. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. URL `http://aclweb.org/anthology/W16-5618`.

Zeerak Talat and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, San Diego, California, June 2016. Association for Computational Linguistics.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, August 2017.

Zeerak Talat, James Thorne, and Joachim Bingel. Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In Jennifer Golbeck, editor, *Online Harassment*, Human–Computer Interaction Series, pages 29–55. Springer International Publishing, Cham, 2018. ISBN 9783319785837. doi: 10.1007/978-3-319-78583-7_3. URL `https://doi.org/10.1007/978-3-319-78583-7_3`.

Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc., 2019.

Yla R. Tausczik and James W. Pennebaker. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1): 24–54, March 2010. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X09351676. URL `http://journals.sagepub.com/doi/10.1177/0261927X09351676`.

Jonas Teuwen and Nikita Moriakov. Convolutional neural networks. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 481–501. Elsevier, 2020. ISBN 978-0-12-816176-0. doi: 10.1016/B978-0-12-816176-0.00025-9. URL `https://linkinghub.elsevier.com/retrieve/pii/B9780128161760000259`.

The Bundestag. Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken. Technical report, German Bundestag, 2017.

Nanna Bonde Thylstrup. Data out of place: Toxic traces and the politics of recycling. *Big Data & Society*, 6(2):205395171987547, July 2019. ISSN 2053-9517, 2053-9517. doi: 10.1177/2053951719875479. URL `http://journals.sagepub.com/doi/10.1177/2053951719875479`.

Nicole Torres. Why Do So Few Women Edit Wikipedia? *Harvard Business Review*, February 2016. URL `https://hbr.org/2016/06/why-do-so-few-women-edit-wikipedia`.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. Detecting racism in dutch social media posts, 2015/12/18 2015.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680. INCOMA Ltd. Shoumen, BULGARIA, 2015. URL `http://aclweb.org/anthology/R15-1086`.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1334. URL `https://www.aclweb.org/anthology/D18-1334`.

Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300, December 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0243300. URL https://dx.plos.org/10.1371/journal.pone.0243300.

Bertie Vidgen, Helen Margetts, and Alex Harris. *How much online abuse is there? A systematic review of evidence for the UK.* The Alan Turing Institute, London, 2019. URL https://www.turing.ac.uk/people/programme-directors/helen-margetts. Backup Publisher: The Alan Turing Institute.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Talat, Austin Botelho, Matthew Hall, and Rebekah Tromble. Detecting East Asian Prejudice on Social Media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.19. URL https://www.aclweb.org/anthology/2020.alw-1.19.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. *arXiv:2012.15761 [cs]*, December 2020b. URL http://arxiv.org/abs/2012.15761. arXiv: 2012.15761.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1702413114. URL http://www.pnas.org/content/114/25/6521.

Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. Detect All Abuse! Toward Universal Abusive Language Detection Models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.560. URL https://www.aclweb.org/anthology/2020.coling-main.560.

William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390374.2390377.

Robert G. Weisbord. The King, the Cardinal and the Pope: Leopold II's genocide in the Congo and the Vatican. *Journal of Genocide Research*, 5(1):35–45, March 2003. ISSN 1462-3528, 1469-9494. doi: 10.1080/14623520305651. URL http://www.tandfonline.com/doi/abs/10.1080/14623520305651.

Sarah Myers West, Meredith Whittaker, and Kate Crawford. Discriminating systems: Gender, race and power in ai, 2019. Retrieved from https://ainowinstitute.org/discriminatingsystems.html.

Risa Whitson. Painting pictures of ourselves: Researcher subjectivity in the practice of feminist reflexivity. *The Professional Geographer*, 69(2):299–306, Apr 2017. ISSN 0033-0124, 1467-9272. doi: 10.1080/00330124.2016.1208510.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North*, pages 602–608, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL http://aclweb.org/anthology/N19-1060.

Hans Friedrich Witschel and Chris Biemann. Rigorous dimensionality reduction through linguistically motivated feature selection for text categorization. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, page 210–217. University of Joensuu, Finland, May 2006. URL https://aclanthology.org/W05-1729.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914, 2016. URL http://arxiv.org/abs/1610.08914.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052591. URL https://doi.org/10.1145/3038912.3052591.

Naomi Zack. *White Privilege and Black Rights: The Injustice of U.S. Police Racial Profiling and Homicide*. Rowman & Littlefield, 2015.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association of Computational Linguistics, 2017.

Jieyu Zhao, Subhabrata Mukherjee, saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.260.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. Improving Hate Speech Detection with Deep Learning Ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1404.