University of Sheffield

Argumentation mining in short text: detecting argumentative information in real-life settings

Anastasios Lytos

A report submitted in partial fulfilment of the requirements
for the degree of philosophy in Computer Science

in the

Department of Computer Science

October 2021

Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: Anastasios Lytos

Signature:

Date: October 2021

Acknowledgements

I would like to thank my supervisors for supporting my efforts the last four years: Thomas Lagkas, Panagiotis Sarigiannidis, Nikos Aletras, and George Eleutherakis.

I also want to thank the organization of SEERC (South-East European Research Centre) for providing my the opportunity to follow my studies in the highest level. I really enjoyed my time there and created amazing friendships with the rest of the PhD students and the staff.

Publications

Part of this dissertation has been published or it is expected to be published in the following months. The list of publications follow:

- Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, Kalina Bontcheva, The evolution of argumentation mining: From models to social media and emerging tools, Information Processing & Management, Volume 56, Issue 6, 2019, 102055, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.102055,
- Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva, 2018, Argumentation Mining: Exploiting Multiple Sources and Background Knowledge, '12th South East European Doctoral Student Conference (DSC2018)
- Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, Vasileios Argyrioucan, George Eleftherakis, Modelling Argumentation in Short Text: a Case of Social Media Debate, Simulation Modelling Practice and Theory (accepted October 2021)
- Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, George Eleftherakis, Nikolaos Aletras, Transfer knowledge to enhance argumentation mining: the use of BERT in real-life settings, 2021 (submitted)

Abstract

The recent technological leaps of artificial intelligence (AI) and the rise of machine learning (ML) have triggered significant progress in a plethora of natural language processing (NLP) tasks. One of these tasks is argumentation mining which has received significant interest in recent years and is regarded as a key domain for future decision-making systems, information retrieval mechanisms, and natural language understanding problems. In the meantime, the developments beyond Web 2.0 have transformed the means of communication and information exchange, promoting shorter bursts of text without solid argumentation. Modern research questions and challenges indicate a need to develop innovative research methods and mechanisms that enhance the trust and the explainability of the AI services enabling the transferability of knowledge between tasks and domains not only on carefully constructed datasets but also in real-life settings.

The main objective of this thesis is to provide a better understanding of natural language in realistic scenarios enhancing the trust in the NLP systems through the task of argumentation detection. Detecting argumentative segments in short text is a crucial step towards a deeper understanding of human language because it delves deeper into the reasoning process and quantifies previously unexplored qualitative aspects. The integration of qualitative aspects through prior knowledge into NLP pipelines has the potential to offer a new perspective and increase the trust in the outcome of AI solutions. This thesis reviews the task of argumentation detection, defines the theoretical foundations for agile argumentation frameworks, offers an annotated dataset to the research community, presents the benefits of integrating symbolic AI into hybrid solutions, and examines the suitability of contextual embedding for the task of argumentation detection.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of natural language processing (NLP) aims to facilitate human-to-machine communication and interaction through the successful modelling of human language. Even though NLP exists as a scientific field for years, the constantly increasing volume of unstructured data generated every day, from medical records to social media posts, provides previously unknown areas for exploration and sets new research questions and goals. Highly unstructured data sources have to be re-organised and decluttered to provide useful textual content and derive new information. The need for exploiting textual data has introduced a series of commercial applications such as smart information retrieval systems, social media analytics, and automated content categorization software. Despite the impressive commercial success, current state-of-the-art approaches, such as deep learning, ignore the need for explainability creating a lack of trust in the latest advances. The need for explainability in NLP tasks underlines the importance of argumentation mining (AM) and the changes in argumentation strategies, due to the extensive use of social media, stress the importance of the argumentation detection task.

The rest of the chapter is organised as follows. An overview of the AM field and the importance of the argumentation detection task in the era of Web-generated text is provided in section 1.1. In section 1.2, the motivation for this thesis is provided stressing the need for adopting novel approaches due to the recent changes in information diffusion channels. Section 1.3 provides the aim of the thesis and presents the research questions that drove this research. The latest advances in the NLP and AM fields are presented in section 1.4 providing the wider context. The main contributions of the thesis are stated in section 1.5, and, finally, the outline of the thesis is provided in section 1.6.

1.1 Overview

The recent advances in artificial intelligence (AI) have created numerous opportunities in many computational areas, including the field of NLP which includes a plethora of higher-level market applications such as machine translation, text summarization, and question answering chatbots. NLP-related tasks include opinion mining, sentiment analysis, and information retrieval, however, the performance of these tasks seems to have reached a plateau [74]. The research questions that emerge are now more complex and require an in-depth understanding of the human language. The progress in the domain provides an opportunity

to improve the understanding of human language, which is a great leap towards knowledge generation. Additionally, the growing demand for enhancing the trust and the explainability of the AI services has dictated the design and adoption of modelling schemes to increase the confidence in the outcomes of the AI solutions. Therefore, it is important to detect and evaluate argumentation, not only in persuasive essays and legal documents but also in short and noisy chunks of text. Noise that occurs due to typographic errors or colloquialisms is evident in online resources, like social media posts and instant messaging. The tendency to use shorter messages and the need for rapid communication often sacrifices semantic clarity and lowers data quality, hence creating what is called noisy text.

As a scientific field, argumentation has been studied extensively by philosophers and computational linguists, focusing on specific aspects such as completeness, effectiveness, and persuasiveness. The initial approaches to modelling arguments aimed at identifying a flawless argument in specific fields (legal text, scientific papers) and serving specific needs (completeness, effectiveness). Until recently, natural language modelling tasks, such as computational argumentation schemes, were often tested in controlled environments, such as persuasive essays. Predefined settings reduce unexpected behaviours that could occur in real-life settings, like a public debate on social media. With the emergence of Web 2.0 and the ever-increasing use of social media, both the diffusion of information and the argument structure have changed drastically. Modern research challenges, such as intelligent information retrieval systems, have brought into the spotlight the need of understanding not only what people think, but also why people take a stance on a given issue. This research question has led to the creation of the AM field which focuses on extracting arguments from natural text, often from unstructured or noisy sources, and performing the tasks of argument detection, relations identification, stance detection, and reliability-related tasks.

The field of AM is defined as a series of actions that could be independent or interconnected and they are relevant to the tasks of detection, extraction, and evaluation of arguments. In the context of this thesis, an argument is defined as a piece of text offering evidence or reasoning in favour or against a specific topic. Similarly, argumentation is defined as the trait of a statement that indicates the existence of an argument. More information on formal definitions can be found in section *3.4.2 Argument quantification* where all the related definitions are presented and illustrated.

The first essential step of every AM pipeline is the detection of argumentation, a task that is often ignored when debate scenarios or persuasive essays are assessed. However, it is extremely important when it addresses real-life settings, such as public debates on social media where arguments, facts, opinions, and fake news coexist. The outcome of this process may be useful in different scenarios, such as advanced information retrieval on a great volume of data, especially on a controversial topic that may trigger lots of discussion and engagement on social media. Argumentation detection can reveal previously unknown useful qualitative aspects to the language enhancing aspects such as transparency and explainability in NLP systems.

The developments beyond Web 2.0 have eliminated the boundaries between the physical and digital world. In the meantime, social media have transformed the means of communication and information exchange, promoting shorter bursts of text without solid argumentation. This new reality calls for an efficient way of detecting argumentative segments that could determine public debates or prevent the spread of rumours and fake news. Despite the extensive

use of social media platforms and their impact on socio-political debates, arguments on social media are often implicit, without a solid logical structure. These settings pose additional challenges that should be recognised, categorised, and solved. This thesis presents novel approaches to face these challenges focusing on the task of argumentation detection in short text. First, a theoretical framework for detecting argumentation is offered, then rule-based and ML implementations are tested, a hybrid solution is evaluated, and, finally, the task of transferring knowledge between different tasks using a pre-trained language model is explored.

1.2 Motivation

The new challenges that have arisen require a deeper understanding of the human language to gain a series of insights into new domains and related tasks. Additionally, the rise of Web 2.0 and the extensive use of social media have created new means and codes of communication disrupting the traditional ones, such as newspapers, journals, and books. The recent shift in shorter bursts of text as a prevalent form of written language has led to structural changes in the use of language [146, 68]. There is often a lack of solid argumentation dictating changes in the argumentation strategies and schemes even in formal contexts [166, 118].

Arguments in informal discourse rarely follow the logical structure of an argument having claims supported by facts, warrants, and qualifiers [58]. On the contrary, they are often implicit without a solid logical structure [137], thus more agile approaches have to be followed for detecting argumentation in a statement and further analysing it. For example, an argument expressed through a tweet is typically a one-sentence argument expressing a stance supported by an external resource without requiring any fact-checking. A successful modelling framework on argumentation in short text would assist a series of NLP related tasks providing a different perspective to established tasks.

During the last decade, there has also been a shift from knowledge- and logic-based approaches towards data-driven solutions, mostly due to the recent advances in machine learning (ML). However, due to a variety of problems in the wider area of NLP, such as hardware limitations and lack of annotated datasets, solely data-driven solutions seem to have reached the upper boundaries of their potential. At the same time, rule-based approaches have remained heavily reliant on domain knowledge facing great difficulty in transferring the gained knowledge to different domains.

In light of these events, it is becoming imperative to develop trustworthy AI applications that can take advantage of the benefits current technologies offer. Towards that direction, improvement in transparency of AI methods is an essential step and should be prioritized. The behaviour of AI models should be explainable and their performance should be trusted. Instead of blindly accepting the outcome of ML algorithms, the aim is to increase the intelligence of NLP systems by capitalizing on the inherited knowledge and eventually, successfully adapting to new situations. It requires the integration of both learning and reasoning on hybrid approaches, which combine data-driven and knowledge-based models and, thus, enable the contextualization of information and the successful discovery of new knowledge. Novel approaches should be based on solid foundations using knowledge modelling methods ensuring their scalability and the re-usability in different domains.

1.3 Aim & objectives

Detecting and breaking down arguments expressed in short chunks of text extracted from the Web or casual discourses is an extremely demanding task. Whether a feasible goal or not, it is accompanied by a series of unprecedented challenges and new research questions. This thesis provides the mechanisms which will advance our understanding of human language through the detection of argumentation in short segments of text. Both theoretical and practical aspects are covered providing a complete overview of the domain. For the needs of the thesis, the task has been narrowed down into smaller parts along with the following research questions.

1. **How is AM connected to modern research and applications?**

Over the past decade, there has been an upsurge in the NLP domain, mostly due to the recent advances in ML. The introduction of ML libraries has allowed the use of ML models off-the-shelf without requiring technical expertise in the domain. Therefore, non-expert users have managed to employ state-of-the-art solutions while commercial applications, such as machine translation and chatbots have been developed and have produced impressive results. However, the frequent need for deep human language comprehension makes the role of AM in research projects crucial. In the meantime, the developments in AM have the potential to spark new commercial applications, such as the smart evaluation of statements on online debates, the development of a badge system in a gamified environment, reliability enhancement of online sources, etc.

2. **How has the rise of social media affected argumentation schemes?**

The structure and rules of natural language have changed moving towards a more informal or casual way of expressing, affected by the trend of social media. Since natural language moves towards shorter chunks of text, traditional argumentation schemes fail to meet these changes and the need for new agile frameworks for argumentation detection is evident. Social media provide ample opportunities for research in the field, as political and social debates attract the interest of the audience. In line with this discussion, it can be argued that argumentation detection in short text is a feasible task with high research and social value, yet a challenging one. The noisy nature of social media texts with the dominance of non-argumentative clauses is the greatest obstacle for traditional argumentation schemes. Future argumentation frameworks should deploy new methods and techniques from related fields that can offer new, valuable insights into the domain.

3. **Have solely data-driven approaches reached a plateau in their performance?**

Over the last decade, the majority of the research community has opted for data-driven solutions over knowledge-based approaches, mostly due to the recent advances in ML. Solely data-driven solutions seem to have reached their peak revealing the need for developing architectures capable of exploiting background knowledge and other formats of data such as audio features and social media metadata. Additionally, aspects such as topic and relation could be integrated into NLP solutions offering useful contextual information. Even though algorithm' optimization and feature engineering is an endless process, traditional ML algorithms do not seem to be able to catch up with the progress of innovative approaches such as contextual embeddings. In this thesis, knowledge-based and ML algorithms are deployed and compared while the benefits of a hybrid methodology are also underlined. Based on its findings, the integration of background knowledge has demonstrated promising results in real-life settings where negatively imbalanced datasets are quite common.

4. **Is it feasible to transfer knowledge from other domains into argumentation detection?**

Dataset annotation is a labour-intensive and expensive process that requires trained experts hence high-quality annotated datasets are in scarcity, especially if we consider the range of the different tasks in the wider NLP domain. In the last years, the NLP research community has shown an increasing interest in methods that can transfer knowledge in different domains by exploiting previous knowledge and enriching it through topic and context learning. Modern NLP systems need to be capable of transferring knowledge and expanding their application fields in new domains. For example, generic language models trained on big datasets can be used for specific downstream tasks such as argumentation detection and stance detection. The argumentation techniques in different topics and different means differ significantly, however, modern AM systems must develop a deep semantic understanding of different topics in different contexts.

5. **What are the social implications of argumentation detection?**

The act of argumentation takes place in our effort to impart our views or analyse and break down the premises in the arguments of others. This explains why computational argumentation initially focused on rhetorics, academic text, and political debates. In the meantime, the developments beyond Web 2.0 have eliminated the boundaries between the physical and digital world while social media dictate the political agenda. In this environment, new technical challenges with strong social implications have emerged such as fake news and hate speech detection. The integration of argumentation detection into more complex applications could provide a greater level of comprehension and present a strong social contribution. For example, social media trends can be assessed not only based on their growth rate but also on -previously unknown- qualitative aspects.

1.4 State of the art and beyond

The wider field of AM has been flourishing in recent years both by researching the field per se and by integrating AM-related tasks into (greater) NLP pipelines, marking its nature as an interdisciplinary field. Despite this surge, there is a lack of formal definitions in this emerging field. The traditional argumentation schemes and definitions fail to surpass the limitations that are imposed by the extensive use of short text segments. For example, in Toulmin's model [174], which still has a great impact on modern argumentation schemes, a detailed microstructure is proposed with six specified components: (1) an indisputable datum, (2) a subjective claim on the foundation of datum, (3) a warrant that links them imposed by logical inference, (4) the backing of the justification which leads to (5) a degree of confidence (qualifiers) as long as (6) a rebuttal cannot withstand the claim. Data that do not have a structured, well-specified format are hard to be represented by such a strict model. Therefore, there is an imperative need for new definitions and frameworks that can be deployed in real-life settings where the structure of the arguments is not clear and implicit arguments dominate public debates.

The extensive use of data-driven solutions has allowed a series of non-experts to execute a series of experiments on a series of computational fields, but it has also created a distorted perception of the field. Machine learning, and AI in general, should not be used as black-boxes but the new approaches should include, among others, the integration of both learning and a

combination of data-driven and knowledge-based models to enable higher levels of causality and contextualization. Many NLP models fail to grasp the wider context and, thus, they are susceptible to biases. Contextual information can improve the transparency of AM pipelines by focusing on the aspects of transparency and explainability. In this thesis, the integration of symbolic AI assists in the explainability of the results, increasing the transparency of the methodology while it provides promising results for the task of argumentation detection.

The contextualization of information is of crucial importance in advanced NLP tasks where a deeper understanding of the human language and reasoning is required. The interpretation of an argument is a natural process that is realized automatically by taking into consideration the wider context and analyzing the different aspects that are related to the discussing topic. However, the modelling and contextualization of text segments is a difficult task since popular encoding methods, such as continuous bag-of-words, do not take into consideration the context of the word but even in the case of methods that do so, such as n-gram, they do not have mechanisms that adjust the weights according to the context. Therefore, words are always mapped to the same vector failing to identify the different meanings a word might have, based on a specific context. As a result, their use in a cross-domain environment is questionable. Contextual embeddings that have been recently introduced in NLP tasks [131, 33] could offer a viable alternative and could be deployed in the field of AM effectively. They seem to have the ability to generalise well and adjust to the task at hand, without requiring a great number of resources in the testing phase.

1.5 Main contributions

The main objective of the thesis is to study and determine the best alternatives for detecting argumentation in short text while providing transparency in the system by increasing aspects such as explainability and trust towards the outcome of the AI algorithms. The contribution of the thesis can be summarized in six points.

1. Frames the problem of argumentation detection in short text, presents a computational definition of argumentation and other related terms in short text, and proposes an agile framework that can be deployed in real-life settings such as social media posts and transcripts of live conversations.

2. Explores the argumentation techniques in Twitter, then proceeds in the annotation of a new dataset for the task of argumentation detection, and finally provides a publicly available annotated dataset on the public debate on the construction of the natural gas pipeline Nord Stream 2, a topic that apart from its economic impact, has also emerged as a political debate in the European Union (EU).

3. Demonstrates a proof-of-concept implementation for the provided framework definition of argumentation in short text utilizing the concept of background knowledge and comparing it with rule-based alternative methods such as sentiment dictionary and Jaccard similarity.

4. Evaluates the performance of four different ML algorithms and the effect of additional features presented through experiment evaluations. The results are compared to pub-

lished findings in real-world datasets providing a clear overview and an in-depth comparison.

5. Implements a hybrid approach that integrates domain knowledge into ML algorithms, confirming the consensus that domain knowledge is beneficial for the task of argumentation detection independently from the source of data. The proposed hybrid architecture outperforms alternative hybrid methodologies that incorporate audio information.

6. Examines the applicability of contextual embeddings for the task of argumentation detection and their feasibility to transfer knowledge to new tasks and domains. The BERT architecture is studied and fine-tuned to better fit the needs for argumentation detection in short text.

1.6 Outline

The overall structure of the study takes the form of eight chapters. Chapter 2 provides an overview of the wider field of AM by presenting the latest related work. Chapter 3 offers the necessary background on both theoretical and technical levels presenting the theoretical foundations of argumentation theory and modelling and presents a novel argumentation framework tailored for noisy environments under the name abstract framework for argumentation detection (AFAD). Chapter 4 provides an overview of the existing datasets in the field, illustrates the best practices for annotating datasets, and presents the process for creating a new dataset. Chapter 5 covers the first case study of the thesis which was applied to social media. It provides different implementations of the AFAD, deploys existing ML algorithms, and proposes a hybrid methodology for the task of argumentation detection. Chapter 6 analyses, and verifies the benefits of the proposed hybrid methodology and presents the second case study of the thesis applied to transcripts from the 2019 UK presidential debate. Chapter 7 examines the concept of transfer learning and assesses the capabilities of contextual embeddings and their successful application on the task of argumentation detection; it fine-tunes the BERT framework and integrates it into a hybrid methodology. Chapter 8 concludes the work of the thesis, evaluates the achievement of goals, and finally presents the future goals of this work. Appendix A provides an overview of the existing tools in the field, presenting tools for different NLP tasks and specialised tools for argument search, retrieval, and annotation.

Chapter 2

Literature review

The recent growth in the NLP domain has triggered the research interest in different tasks, such as machine translation, sentiment analysis, and speech recognition. The field of argumentation mining (AM) has also emerged trying to gain a deeper understanding of human language and reasoning. Under the field of argumentation mining, there is a series of interrelated distinctive tasks, such as argumentation detection, with unique characteristics and different challenges. In the era of social media where solid arguments, personal reflections, and fake news co-exist, the task of argumentation detection can reveal some previously unexplored qualitative aspects. This chapter covers the latest related work in the field of AM presenting its connection with opinion mining, exploring in-depth the task of argumentation detection, and examining every task in its wider context.

The rest of this chapter is organised as follows. In section 2.1 the related work in the wider field of AM, is presented illustrating the evolution of the field. Section 2.2 presents extensively the related work for the task of argumentation detection. The following three sections cover tasks that are often included in AM pipelines; section 2.3 is dedicated to the task of relations identification in argumentation schemes, section 2.4 presents the task of stance detection, and section 2.5 presents reliability-related tasks, such as evidence classification and source identification. Finally, section 2.6 initiates a discussion based on the findings of the literature review.

2.1 From opinion to argumentation mining

The research questions that have emerged in the last decade created different tasks in the NLP domain. For example, opinion mining answers the question of what people think on a given topic, sentiment analysis interprets and quantifies human feelings, and dialogue management mimics the human conversation successfully replacing human assistance in trivial interactions. The field of AM aims to gain a deeper understanding of the human language by answering why people hold a specific view towards a topic. It is the evolution of computation argumentation while it is also influenced by the task of opinion mining integrating aspects such as argument and stance detection. Opinion mining and sentiment analysis could be characterized as the predecessors of AM in a simplified form and their limits have already been questioned in the effort of seeking a deeper understanding of the human reasoning [9, 175].

The field of AM is identified as a multidisciplinary research topic having its roots in

rhetoric and philosophy [174], and it recently gained the interest of the scientific community due to the progress in the field of AI. The recent advances in ML and the emergence of the social Web have enabled impressive progress in different scientific fields with a great impact on commercial applications. For example, machine translation applications and writing assistants present impressive results and have great commercial success. It is not expected from an AM system to produce a stand-alone application, but its outcome is expected to create a feature for an existing application, similar to the classification of product reviews based on the findings of the opinion mining task. An AM system can mine and analyse a great volume of text data through a variety of sources, providing tools for policy-making and socio-political sciences [92, 1, 14], and software engineering [82]. Additionally, it opens new horizons for the broader area of business, economics and finance, with digital marketing being the most promising field [120, 121, 136, 138].

Tasks that are included under the term AM attempt to solve a series of problems such as the detection of an argumentative stance towards a specific object, the analysis and evaluation of the argument's components, and the detection of possible relations between them. The cohesion in the components of the arguments and the existence of backing in the claims is a major challenge in AM because they can alter a disputable claim into a valid argument. During a written or verbal discourse the interpretation of an argument is often realized instantly by its participants without requiring any special effort. Their skill to grasp the context of the information and to accomplish connections with previous experiences (facts, opinions, feelings) assist in the completion of tacit assumptions or premises (enthymemes) avoiding logical impasses.

This ability to combine multiple sources of information is missing in existing argumentation models as it is difficult for human annotators to agree upon specific guidelines for the modelling of an argument. Furthermore, identifying the thin line between implicit and explicit arguments is harsh as often annotators are implicitly driven by the context. Probably that is also the reason why the majority of research was initially focused on structured data like law text [109, 150, 93], scientific text [55, 86, 87], formal debates [113] or news articles [6, 149] instead of unstructured text like informal discourse and Web-generated data. However, recently there have been some notable research endeavours in this direction.

Annotating and automatically analysing arguments from the Web with great heterogeneity of content and diversity of jargon is a challenging task. Arguments in social media and informal discourse are sometimes implicit, meaning that the logical structure of an argument's components (premises, claims, warrants, etc.) are not always spelt out and instantly distinguishable hence an in-depth analysis must take place to determine the distinctive components. For example, it is common a tweet or a Facebook post to contain just a stance on a specific topic without supporting it with evidence or reasoning.

The decoding of the human reasoning process into computer language is a challenging task because it consists of many subprocesses that are difficult to be separated and analysed. The medium for arguing for human beings is natural language, whereas the input for ML algorithms and techniques should be distinct, structured and composed of well-established rules. A wide range of methodologies have been applied to modeling natural language, such as explicit distinctive components [174], argumentative zoning [141], tree structures, dialog-oriented diagrams [155], serial structure of arguments [44, 182] and modifications to simpler structures of existing schemes [126, 160].

However, the claim or other parts of an argument might be implicit [54, 136, 15, 60, 138] and enthymemes take place related to commonsense reasoning. This process is named completion or enthymematic argumentation and takes place often and unconsciously in casual discourse that can be found in Web-generated data. The distinction between explicit and implicit arguments lies in the presence of certain syntactic constructions or lexemes (such as conjunctions). Implicit arguments, where the lack of these characteristics is noticed, can be identified through previously gained knowledge and logical inference. This a priori knowledge is extremely difficult to be expressed through conventional argumentation schemes, which demand a strict structure of the components of the argument. The early approaches [174, 103] were focusing on the philosophical aspect of the argument, whereas modern approaches consider unstructured data and implicit relations between the components of the argument [54, 136, 15, 60, 138].

In the era of Web 2.0, breaking down arguments deriving from the Web or casual discourse is a demanding task with doubts, if it is even feasible. Evidence of the previous statement is the fact that the field of opinion mining thrives on social media data [90, 115] and especially in Twitter [24, 49]. On the contrary, only limited research has been conducted on AM in unstructured data and fewer frameworks have been designed which are able to capture the special features of social media. Future argumentation modelling schemes should seek novel approaches to the problem of detecting argumentation in real-life settings and mapping qualitative characteristics to quantitative features. They should be more agile providing the necessary flexibility to researchers to implement different versions in different contexts.

2.2 Argumentation detection

In the noisy environment of social media, it is important to develop mechanisms that are capable of identifying argumentation in short text revealing previously unexplored capabilities in the wider spectrum of the NLP domain. Short text, and especially Web-generated text, like the one produced on social media, does not have the format that is required to apply traditional argumentation schemes. It becomes clear if an attempt is made to apply Toulmin's scheme to a typical tweet. According to Toulmin's scheme, an argument consists of six different pieces: grounds, backing, warrant, qualifier, claim, and rebuttal; parts that are hard to find on shorter online statements. Traditional argumentation schemes were designed to be applied to texts of legal nature with an excellent reasoning process, and not on tweets or Facebook posts. More information on logical schemes and diagrams is presented in the following chapter, in the section *3.1 Logical schemes and diagrams*.

On the other hand, in the work of Addawood and Bashir [1] the following tweet from their dataset is presented: *RT @ItIsAMovement "Without strong encryption, you will be spied on systematically by lots of people" - Whitfield Diffie*. The above tweet cannot easily fit in Toulmin's scheme (or any other theoretical scheme) because the fact is the same as the conclusion of the argument, and the backing is expressed through the quote of an expert opinion. A similar belief is also expressed in Bosc et al. [18] where the authors claim *"we (almost) never find such a kind of complete structure of the arguments"*. It becomes clear that there is a need for adopting agile approaches for the task of argumentation detection capable of capturing argumentation in short and noisy text.

Argumentation detection gains a prominent role when AM pipeline is deployed on Web-

generated data where the structure and completeness of the arguments are undervalued for the sake of speed and immediacy. In this thesis, argumentation detection is defined as a binary classification task that evaluates the existence or not of argumentative elements in a given statement. Even though argumentation detection is frequently neglected in some NLP tasks, it is an integral part of human reasoning. The nature and peculiarity of social media data underline the importance of the task of argumentation detection because the public speech on these means is often characterized by high polarity and a lack of solid arguments.

2.2.1 Transforming unstructured data to quantitative features

Algorithms do not receive as input the human language per se, thus it is crucial to extract the most useful features from the textual input. The extracted features should express the content of the input text creating a model that can decipher the arithmetic input into the actual meaning. Some of the common features that are used in the NLP domain include the number of words in a sentence, the size of a sentence, the existence of different parts of speech, etc. With the advent of Web 2.0, the online text has been enriched with meta-data that can be used as additional features providing another perspective on the input text. Additionally, multi-media features, such as audio fluctuation, can also enhance the performance of the NLP systems if the former are properly aggregated.

Authors	Basic (Lexical)	Semantic	Sentiment	Subjectivity	Twitter	Other
Addawood and Bashir [1]	n-gram, length, question, exclamation	PoS, LIWC summary, variables	LIWC, sentiment, lexicon	Clue lexicon	followers, friends, user activity, URL+title, hashtags, verified account, mentions	Psycometric, LIWC
Bosc et al. [18]	n-gram, punctuation, tokens, capitalization	PoS			smileys	
Deturck et al. [32]	tokens, lemmas	PoS	TextBlob features	TextBlob features		word embeddings
Dusmanu et al. [36]	n-gram	Syntactic tree, parse trees, dependency relations WordNet synset	AlchemyAPI		Punctuations, emoticons	
Sendi and Latiri [154]	IR, TM		NRC emotion lexicon			
Dufour et al. [34]	punctuation,	personal pronouns	Emotion words		Emoticons, hashtags	

Table 2.1: The different features used for the argumentation detection task. The explanation of the acronyms used in the table follows. IR = Information Retrieval applied as lexicon-based queries, PoS = Part of Speech, TM = Topic Modelling applied with the technique of Latent Dirichlet Allocation (LDA), LIWC = Linguistic Enquiry and Word Count

Table 2.1 summarizes the features that are used in related work for the task of argumentation detection summarized in six different categories. The lexical features are the attributes that are used most frequently in the wider spectrum of the NLP and they are often correlated with the different applications of n-grams. The semantic features are the characteristics of the language that can provide a deeper insight into the data. The sentiment features are those

which can trigger emotions and usually they are detected with the use of specific lexicons or libraries. The subjectivity features often indicate an opinionated and therefore an argumentative tweet. The twitter-specified features are offered as metadata through the Twitter API and are related to the specific characteristics a tweet contains, such as its length or the number of hashtags it contains. In the last column are the features that cannot be grouped under any of the previous categories. Apart from the semantic and sentiment features, LIWC offers statistics that include personal concerns, core drives and needs, which are summarized under the psychometric category [128, 127]. In the work of Deturck et al. [32] the use of word embeddings takes place, aiming at a diversity filtering for the most argumentative tweets to be discovered. Overall, the number of features that are used in the ML algorithms does not seem to pose additional computational constraints because the resource-demanding training phase is executed only once. Additionally, real-time execution on heavy volume datasets does not represent real-life scenarios.

It must be stated that the classification of features is different in Addawood and Bashir [1], where classifying emotional tone and subjectivity score under the linguistics features. Furthermore, the tasks of information retrieval and topic modelling are included as lexical features [154], provided that they do not have any semantic purpose. Those adjustments were made to provide a useful taxonomy and a beneficial comparison, but it should be noted that different categorizations can take place.

2.2.2 Performance of current approaches

Having provided the necessary introductory information on argumentation detection and feature engineering, this sub-section presents the performance of current approaches. The performance of the different solutions for the task of argumentation detection varies depending on the dataset, the features that are used, and the actual algorithms that are deployed. Only supervised algorithms can be evaluated while the unsupervised ones are usually subject to qualitative analysis. Although the one-to-one comparison between different research studies is often not reliable because other factors should also be included, such as the quality of the dataset, they still offer a benchmark for comparison. It should be noted that there are not any rule-based methodologies that have been applied on the task of argumentation detection in short text, at least to my knowledge.

The evaluation of the algorithm's performance is usually achieved using three different metrics (precision, recall, f1-score), each one presenting a different aspect of the algorithm's behaviour. Precision is the ability of the classifier to not label as positive a sample that is negative. Recall, on the other hand, is the ability of the classifier to find all the positive samples. Finally, the weighted average of precision and recall is the F1- score and it is considered the most reliable method in the cases of balanced datasets. Providing a wider perspective, precision is the ability of the classifier not to label as positive a sample that is negative and when the priority is the detection of positive instances on a negatively imbalanced dataset, the importance of precision declines. Recall, on the other side, is the ability of the classifier to find all the positive samples and in many real-life scenarios when the positive class is the minority, the goal is to increase the recall because it discovers more positive instances, which often is the requirement. Finally, the F1-score is considered the most reliable method while its different calculation methods (binary, mico, and macro) can provide an in-depth overview of the performance of the algorithms.

Authors	Algorithm	Features	Precision	Recall	F1
Addawood and Bashir [1]	DT	n-gram	0.72	0.69	0.66
	SVM	n-gram	0.81	0.78	0.77
	NB	n-gram	0.70	0.67	0.64
	DT	all features	0.87	0.87	0.87
	SVM	all features	0.89	0.89	0.89
	NB	all features	0.79	0.79	0.85
Bosc et al. [18]	LR	lexical	-	-	0.64
	LR	lexical + semantic	-	-	0.66
	LR	all features	-	-	0.67
Dusmanu et al. [36]	RF	n-gram	0.76	0.69	0.71
	LR	n-gram	0.76	0.71	0.73
	LR	all features	0.80	0.77	0.78

Table 2.2: The results of the supervised ML algorithms for the task of argument detecton. Un-supervised methods are not included. The explanation of the acronyms used in the table follows. RF = Round Forest, LR = Logistic Regression, DT = Decision Tree, SVM = Support Vector Machine, NB = Naive Bayes. Rounding took place in order for the results to be displayed in the same scale.

Table 2.2 presents the performance of related work on the task of argumentation detection. The first column depicts the names of the authors, whereas the second and the third columns present the algorithms and the features that have been used, respectively. The last three columns show the metrics that have been used to evaluate the performance of the algorithms. Regarding the feature selection and their impact on the classification task, the use of all possible features performs better in each case. The selection of the classification algorithm does not seem to significantly affect the performance of the task in Dusmanu et al. [36], whereas in Addawood and Bashir [1] the use of SVM surpasses the alternative algorithms. The best results are achieved in Addawood and Bashir [1] with 0.89 F1, whereas Dusmanu et al. [36] and Bosc et al. [18] achieve 0.78 F1 and 0.67 F1 respectively. In terms of precision, in Addawood and Bashir [1] the SVM has the best performance when all the features are used reaching 0.89, and in Dusmanu et al. [36] the LR with the use of all the features reaches 0.80. The same algorithms present also the highest score in terms of recall, presenting 0.89 and 0.77, respectively.

The algorithms that are used differ in complexity, but this does not seem to be an issue of major concern since there is a lack of annotated datasets for argumentation detection. The depth of a tree that is created using DT is *O(log(n))*, where n is the number of samples and for every training stage it requires a decision to be taken for every feature/dimension d and every sample d leading to *O(n*d*log(n))* complexity. The SVM has been deployed using the LibLinear [41] which excludes non-linear problemsm lowering the complexity to approximately *O(n)* where n is the number of samples. The NB proceeds in a conditional independence assumption to extend the Bayes theorem, eventually providing an algorithm with the complexity of *O(n*d*c)* with the n representing the number of samples, the d the

dimensions of the dataset, and c the number of classes. The logistic regression deploys a logistic function producing a solution with *O(nd)*, with the n representing the number of samples, the d the dimensions of the dataset. Finally, the RF is an ensemble model of decision trees that requires the complexity to create a DT, *O(n*d*log(n))*, and the number of trees t created leading to *O(n*d*log(n)*t)*.

2.2.3 Alternative and cutting-edge approaches

The majority of the research in the NLP domain, including AM related tasks, has adopted data-driven solutions mostly due to the recent advances in off-the-self ML libraries. However, hybrid solutions are used in different AI applications and combine traditional ML algorithms with domain knowledge present a viable alternative offering a greater degree of explainability and transparency. The integration of a rule-based expert system into ML algorithms for the task of text categorization in the Reuters-21578 dataset improves the F1 performance of the kNN up to 7.3% [178]. A similar approach on 5 different Twitter datasets confirmed the added value of enhancing sub-processes that determine the final class of a tweet [76]. The concept of argument base (AB) is introduced as part of a four-step classification process using a collection of arguments and relations as a middle step to identify if the argument in a text segment attacks or supports the original statement [22]. For the task of claim detection, an innovative approach was followed enhancing the ML algorithms with audio features and eventually increasing the F1-performance of the SVM model up to 8.4% [95].

Apart from integrating rules into traditional ML algorithms, the NLP research community has shown an increasing interest in methods that can transfer knowledge to different domains by exploiting previous knowledge and enriching it through topic and context learning. The BERT architecture, a language model created for machine translation, has been a game-changer for a wide spectrum of NLP tasks [33]. For the task of argumentation detection in the climate change discussion in the German language on Twitter, pre-trained BERT embeddings outperform the bi-gram solution, but on the other hand, for the task of evidence detection, BERT embeddings seem not to perform better [152].

At document level, the information about the topic and the context of the debates, for a broad range of genres, increased the performance for both binary and three-class classification problems [45]. Providing some more technical details, six different recurrent neural network implementations have been tested using contextual information from shallow word embeddings, knowledge graphs and pre-trained transfer learning approaches; with the latter option having been implemented with BERT and surpassing the other methods. With a similar approach being followed for the task of binary claim detection task on online persuasive discussion forums (CMV subreddit), fine-tuned BERT improves by 5 points the pre-trained BERT solution (0.70 F1-score) while it surpasses existing state-of-the-art solutions for the tasks of argument component classification and relation prediction [23].

For the same side (stance) classification task the use of BERT surpasses the SVM baseline, while also confirming notions regarding the benefits of larger transformer models and longer input sequences [117]. On a similar note, the benefits of contextualized word embeddings are highlighted through the integration of topic information into the BERT architecture on two different document-level datasets [143]. The potential of transfer learning to address AM tasks (stance classification, evidence detection, argument quality) in non-English languages using the multilingual BERT (mBERT) model has also been explored [171]. The deployed

method provides insight into the usefulness of machine translation in AM related tasks, indicating that simpler tasks can be assisted in contrast with more refined tasks such as argument quality. Finally, delving deeper into the BERT architecture, a procedure to isolate and combine the "best heads" has been suggested, improving the ability to detect arguments for a given event role [47]. Overall, BERT presents satisfactory results in different results and it offers a new domain for novel experiments which is not yet saturated.

2.3 Relations identification

Another aspect of AM that is also worth researching is the detection of different argumentation models and the identification of relations between the components of an argument. The choice and the customization of the theoretical argumentation model that will be adopted affect the individual tasks that will be raised. Especially the task of relations identification, or argument structure prediction as expressed in Lippi and Torroni [96], is the part of the AM pipeline which is the most susceptible to potential changes in the adopted model. The task of annotating relations between parts of text requires the adoption of a holistic approach, capable of identifying connections with both preceding and succeeding components of the selected model, the relations between the entities of the network and eventually offering a better understanding of the argument.

Both Toulmin's [173] and Freeman's [43] theories, two of the most influential theories in the wider field of logic and argumentation, explicitly define relations between the components of the arguments. In data derived from social media, argument component identification is a challenging task as both their size and their chaotic nature do not allow strict rules and principles to be applied. As a consequence of this situation, the task of relations identification should be redefined and include both micro and macro analysis. The micro-analysis evaluates the quality and the completeness of the argument, whereas the macro-analysis expresses the relation of an argument either towards a known topic or towards an argument previously expressed. In social media, network analysis algorithms can significantly boost macro-analysis tasks, as the introduction of network-based features reveal underlying relations between the users and eventually improve the prediction model [83].

The possible outcomes of a macro-analysis in social media text are limited to support/attack/neither relations indicating the outcome of the stance detection to a great extent. On the other hand, micro-analysis is related more to AM and other reliability-related tasks, as it evaluates the integrity and the cohesion of the argument. The arguments extracted from online resources are not characterized as high-quality data, as often complicated reasoning processes should take place for the argument to be understood. Arguments with missing premises are called enthymemes [136, 138] and take place often in informal discourse, creating a challenge for the approach that should be followed; discard the argument or try to fill the missing premise.

The difference between micro and macro analysis has not been researched in-depth in the field of AM, hence a categorization of the different tasks, and approaches within the tasks, could assist in clarifying the distinction between them. The building block in every research scenario is a chunk of text, which could be on different dimensions, and in the short-text analysis, they are usually either on sentence-level having a sentence as a core, or actor-based focusing on the actions and interactions of the user. The main difference between micro and

macro analysis is that the first one is limited by the sentence's dialectical limits focusing on the information that already exists, while the latter aims at collecting information from external resources and connecting different information segments.

Table 2.3 presents the different tasks and approaches in the literature that fall under the scope of micro and macro analysis. The first column presents the scope (macro or micro) of the analysis, the second column describes the different tasks, and finally, the third column presents the different approaches that have been used to undertake the tasks. The first example of analysis of the macro-level is the task of pair creation between a segment of text (e.g., a tweet) and an argument from a manually constructed list when encountered as a classification problem. The tasks that require the construction of graphs such as the development of network communities and detection of uncrossing stance relations, fall into the macro-level analysis when requiring external resources. On the other hand, if the graphs are built based on a rule-based approach exploiting argumentation semantics within the body of the text are analysed at the micro-level. Similarly, detecting arguable reason could be on the macro-level when external resources are used, or micro-level when a rule-based logic is applied. Finally, the task of relations detection within the given chunk(s) of text, approaching the problem as a textual entailment problem, is an analysis on a micro-level because the process does not require any external resources, whereas a pre-trained neural sequence classifier such as when the GloVe embeddings are used is considered analysis on the macro-level.

Scope	**Task**	**Approach**
Macro	Pairs creation [18]	Classification problem
	Creation of network communities [83]	Graph creation
	Uncross stance relations [83]	Graph creation
	Arguable reason [101]	Lexical patterns
		Use of external lexicon
	Relation detection [27, 18, 101]	Classification problem
		Pre-trained neural sequence classifier
Micro	Graph building [18]	Rule-based
	Relation detection [18]	Textual entailment
	Arguable reason [27, 101]	Rule-based

Table 2.3: Taxonomy of micro and macro analysis for the different tasks for relations identification in the context of AM.

The need for both micro and macro analysis for the task of relations identification, in combination with the low-quality data from social media, creates the need for the establishment of simple, but effective rules and standards. Towards the need for providing a straightforward definition able to capture both micro and macro analysis, we define three entities (argument, topic, completeness) able to capture the complicated nature of the task. We convert the problem to a mathematical expression in a triple, where the task of relations identifications is split into two parts, where the first part connects the argument with a specific problem (favour/against/neither) and the second part evaluates the structure of the argument. Eventually, expressing the identification of the relations as a triple we have:

$$(a_{\text{ij}}, t_i, c_j)$$

where the expressed argument a_{ij} is open to a macro-analysis considering a topic t_i and a micro-analysis for its completeness c_j. The subscripts indicate the relations between the different components, showing that the topic and the claim do not have a direct interaction.

Only a few researchers have explored the task of relations identification in text derived from social media, because of the chaotic nature of social media and the wide presence of vague claims. The first step (argument detection) in the proposed pipeline of Bosc et al. [18] was presented, which is followed by the prediction of attack/support relations between tweets and arguments. Their adopted approach is similar to textual entailment, thus the excitement open platform (EOP) and the recognizing textual entailment (RTE) framework were used. A second method was also applied to implement a neural sequence classifier, however, none of the methods presented satisfactory results. In fact, the detection of support-relation achieved 0.20 F1 and the attack-relation 0.16 F1 using a neural model, and with the use of EOP+RTE, the support-relation achieved 0.17 F1 and the attack-relation 0.0 F1. Apart from the low score in the automatic detection of relations, even the IAA was significantly lower a=0.67 for the specific task, compared to the IAA for the task of argument detection which reached a=0.81.

A new method for extracting argumentative relations of attack, support or neither is presented in [27] based on the Relation-based Argumentation Mining (RbAM) model. The proposed model was tested in the dataset of Carstens and Toni [22] and afterwards, it was applied to the task of relations prediction between tweets and news headlines on two different datasets [57, 168]. Apart from different implementations of neural networks, the impact of trained and non-trained was also evaluated demonstrating the supremacy of the trained embeddings. Apart from the use of word embeddings and the argumentativeness features extracted from RbAM, the authors do not describe the rest of the features, instead, they simply use the term standard features, thus safe conclusions for the use of features cannot be drawn.

Broadening the limits of relations prediction task, Ma et al. [101] introduced a 3-step framework including both micro and macro-analysis of the argument. Besides the attack/support relation between a tweet and a topic, the authors also examined the relatedness of a topic towards the pre-defined topic and the existence or not of an arguable reason, where the evaluation of the argument toward its completeness takes place. Considering the complexity of the proposed methodology and the comparison with state-of-the-art baselines [111, 51, 145], the presented results can be characterized as promising. The entire process is characterized as an information retrieval task, thus the learning-to-rank approach was adopted and the metric precision at k was used, which indicates the precision among the k top results of the retrieval. The three distinctive sub-tasks that make up the retrieval task increase its complexity, as it is pointed out through the report of the IAA, where topic-relevance reached 90.1%, clear stance 78.2% and detection of arguable reason 75.2%.

In Table 2.4, a summarization of research papers on the task of relations identification in text derived from social media takes place. The first column of the table presents the authors of the paper, the second one interprets the scope of the task(s) according to the proposed definition, and the task is presented in the third column. The next three columns present the technical details (algorithms, metric, score) for the implementation of each proposed method. In [27] the results for the two datasets that their proposed methodology has tested are presented. It should be stressed that the scores are not directly comparable, as different

Author	Scope	Task	Algorithms	Metric	Score	
Bosc et al. [18]	macro-analysis	support / attack	EOP + RTE	F1 support	0.17	
				F1 attack	0.0	
			LSTM	F1 support	0.20	
				F1 attack	0.16	
Cocarascu and Toni [27]	macro-analysis	support / attack / neither	LSTM		Dataset 1 [168]	Dataset 2 [57]
				P	0.59	0.97
				R	0.97	0.90
				F1	0.73	0.94
Ma et al. [101]	micro & macro analysis	topic relatedness, support/attack, arguable reason	SVM light	MAP	0.59	0.50
				P@5	0.53	0.51
				P@10	0.48	0.44

Table 2.4: The different sub-tasks that have been accomplished in the context of relations identification task. The explanation of the acronyms used in the table follows. MAP = Mean Average Precision, P@5 = Precision@5, P@10 = Precision@10. Rounding took place in order the results to be displayed in the same scale.

research papers carry out different tasks; instead, we should focus on the coexistence of different approaches and the level of difficulty of each one.

Each one of the presented research papers exploits data derived from Twitter since it is the primary source of freely accessible short text. Web-derived text seems to thrive as a source of AM pipelines, including the task of relation identification, but the source of data usually is a more structured source of text, such as debate forums [89, 112, 46, 39]. Although the information found in social media is characterized as noisy text and it is far from an ideal scenario for AM, the constant generation of content allows the researchers to conduct research including the time axis in order to understand users' behaviour [83, 84] and evaluate their impact beyond the network [107, 29]. Users in social media platforms usually express emotions or quick messages with very little argumentation, however, the introduction of argumentative features can enhance other NLP tasks [2, 27]. Both micro [153] and macro [88] analysis have the attention of the research community, whereas there have been approaches that combine them [112, 46]. Finally, another research topic that has gained the interest of the research community is the reconstruction of implicit warrants, although the existing research papers do not utilize social media as a source [136, 138, 60].

2.4 Stance detection

In contrast to the relations identification task, stance detection is a popular task among researchers in the NLP community, either as an autonomous and independent task or as a part of an extensive pipeline. It is thriving even in the challenging environment of social media and it is related to many sub-fields of the wider NLP area, such as sentiment analysis, textual entailment, and topic extraction. In the context of AM in social media, we define stance detection as the task of automatically determining the attitude of the author towards

a specific topic by exploiting any kind of information that can be collected. The stance can be determined either exclusively by the content of the text either from the combination of features that are capable of revealing specific characteristics such as argumentativeness [2] or network communities [83, 53].

The main difference between opinion mining and stance detection, as it is expressed in AM pipelines, lies in the concept of data aggregation from the wider environment towards the final outcome. Stance detection is considered as the final part of the pipeline that exploits the findings of the previous steps, rather than a stand-alone task. The term wider environment applies to both combination of sources and tasks [99], as Web-generated data and especially social media offer an excellent environment for sentiment analysis, but a poor one for argumentation or opinion mining.

As the research community has shown great interest in the task of stance detection, it is impossible to present each research paper, rather we decided to focus on the research methodologies that either aggregate data from different sources or have the ability to be used as part of a bigger system. Similar to the relations identification task, the mathematical expression of the stance detection is also provided, influenced by [98] on the definition of opinion mining. Stance detection is expressed as the quintuple:

$$(h_i, s_{\text{ijkl}}, d_j, r_k, t_l)$$

where h_i is the person who holds a specific stance s_{ijkl} for a specific debate d_l, justified by a rationale r_k in a specific time t_l. The rationale of the stance for a specific debate can be assessed for its quality through a variety of sub-tasks, such as facts identification, evidence recognition, source classification and reasoning evaluation [36, 1]. Similar to the formula provided for stance, the subscripts express the relations between the different components of the proposed scheme.

The sixth task of SemEval-2016 [110] introduced the shared task of stance detection in tweets, providing a significant boost in the field as new methodologies were suggested and the constructed dataset was also used in later research. The shared task consisted of two parts regarding the supervision framework to be followed (fully-supervised, weakly-supervised). As the constructed dataset was used after the completion of the task, more modern approaches have surpassed the top performances described in the task.

The best-performing system of the competition [191] proposed a recurrent neural network capable of extracting information from unlabeled datasets using word embeddings. The use of word embeddings as features was also critical in [111], where a simpler linear SVM algorithm achieved an F1-score up to 0.70, surpassing the previously highest score of 0.68. Apart from the use of word embeddings, the presence or absence of the target of interest in the tweet improved the results of the algorithm. One more improvement in the same dataset for the same task was achieved by Wei et al. [184] reaching the F1 to 0.71, where an end-to-end neural model was proposed which makes better use of target information.

Considering the results for the weakly-supervised framework as described in [110], those are significantly lower as no training data are provided for the topic that is researched, thus the developed methodologies rely heavily on techniques that can transfer knowledge from different topics. The submission with the highest performance for the task achieved 0.56 F1 [185] and proposed a convolutional neural network including a modified softmax layer able to perform three-class classification, although the training consists of two classes. A weakly-

supervised approach exploiting the network structure information proposed by Ebrahimi et al. [37] improved the previous best score and reached 0.57 F1.

Weakly supervised approaches exploiting social media features for political stance detection were also adopted in Lai et al. [83] and in Johnson and Goldwasser [72]. The former is focused on the Italian referendum in 2016 and employs a holistic approach for stance detection adopting a diachronic perspective of the user's stance including Twitter-specific features and social network communities, achieving 0.90 f-micro with the use of SVM. A similar approach was also followed in [72], able to capture both the content and the social context through linguistic patterns reaching 0.86 accuracy. The novelty in their approach lies in the absence of manual annotation, as the annotation of the political stances took place with the use of ISideWith.com.

In the work of Addawood et al. [2], an advancement of a previously established scheme [1] is proposed as capable of carrying the task of stance detection resulting in a F1 score of 0.93 with the use of the Decision Tree algorithm. In their work, the introduction of argumentativeness as a feature takes place significantly increasing the performance of the algorithm. Their findings indicate that argumentativeness features are the most informative ones for the successful categorization of favour and neutral categories, stressing the importance of introducing AM techniques in different text mining tasks.

Author	**# classes**	**Algorithm**	**Metric**	**Score**
Zarella and Marsh [191]	Favor/against/neither	RNN	F1	0.68
Mohammad et al. [111]	Favor/against/neither	linear-kernel SVM	F-micro	0.70
			F-macro	0.59
Wei et al. [185]	Favor/against/neither	Neural network	F1	0.56
Wei et al. [184]	Favor/against	Neural network	F1	0.71
Ebrahimi et al. [37]	Favor/against/neither	Linear-kernel SVM	F macro	0.57
Johnson and Goldwasser [72]	Favor/against	Probabilistic Soft Logic	A	0.86
Lai et al. [83]	Favor/against	SVM	F-macro	0.90
Addawood et al. [2]	Favor/against/neutral	SVM	P	0.90
			R	0.90
			F1	0.90

Table 2.5: The results of the supervised and weakly-supervised ML approaches that have been followed for the stance detection in social media text. Fully supervision on [191, 111, 2]. Weakly supervision on [184, 37, 72, 83]. [191, 111, 185, 2] are applied on the same dataset. RNN = Recurrent Neural Network

Table 2.5 presents the algorithms, the metrics and the score achieved by selected research papers. In the second column, the possible classes that are provided to the classification algorithms are depicted, and there is either a binary approach (favour, against) or the 'neither' option is included. Considering the algorithms that are used, SVM and neural networks algorithms are used, apart from [72] where a probabilistic soft logic was adopted. In the last two columns, the scores that were achieved measured with different metrics are presented. As expected fully supervised ML approaches achieve higher results when compared to weakly supervised ML approaches when both are applied to the same dataset. Apart from the ML approach, defining the number of possible classes plays a role in the performance of the algorithms, as the binary approach achieves normally higher results.

2.5 Reliability-related tasks

The wide use of Twitter in combination with its public nature has established it as the most appropriate social network for studying a variety of tasks related to virality, such as viral marketing, rumour diffusion, event and fake news detection. The majority of the proposed methods rely on social network analysis, exploiting the metadata offered by the social network (friends, followers, time of publishing, etc.). As the scope of this paper is neither an in-depth review of rumour detection techniques and methods nor the evidence identification for claims in any kind of text, we are going to focus on research work that connects argumentativeness with the evaluation of the argument's reliability in social media text.

Twitter is often used as a means of expressing arguments for controversial topics; some of them are efficiently supported with facts and evidence from reliable sources, whereas in some other cases, instead of backing their claims, they simply express feelings or unsupported allegations. The constant data generation and rapid pace of news flow create a chaotic environment with limited time for claims to be evaluated and facts to be assessed. Due to the environment that has been created, where users express opinions and views in real-time without using sophisticated or pretentious vocabulary, a unique opportunity is raised for argument evaluation on various political and social issues. The automatic evaluation of arguments has the potential to reduce the incidents of rumour spreading, the faster detection of fake news and eventually the improvement of the quality of public political discourse.

An essential part of a complete argument is the sufficient backing of the original claim, either in the form of premises or in the form of backing with evidence and facts presentation. A simple, but robust structure of claim and supporting evidence is adopted in Addawood and Bashir [1] for the classification of arguments' evidence, where the ultimate step of the proposed pipeline is the classification of evidence into six different categories. The proposed pipeline of Dusmanu et al. [36] contains two tasks that are related to the reliability of the expressed argument, the distinction of factual information from opinions and the source identification from a pre-defined list. For both tasks, the use of all the available features boosts the results of the classification.

In the work of Konstantinovskiy et al. [78], where the objective is the construction of a reliable fact-checking mechanism, both an annotation scheme and an automated claim detection method are proposed. Two different approaches are presented, the binary model of claim/no claim and a multi-class classification with seven categories describing the claim. The proposed methodology (the binary model) overcomes previously established mechanisms (Claimbuster, ClaimRank) in terms of F1, as it achieves F1 0.83, while the multi-class classification displays the impressive 0.70 F1-micro, and in terms of macro average, it achieves 0.48 F1-macro.

In Table 2.6 a summarization of the research work that accomplishes reliability-related tasks in the context of AM in social media text is presented. Four tasks have been recognized in this category and the results of the different approaches heavily rely on the number of the alternative classes. The 0.83 F1-macro that is achieved in [1] is impressive if we consider that there are six available classes, and for a similar task with seven available classes the F1-macro is 0.48 in [78]. The difference could be merely explained through the exploitation of more features in [1] in comparison to [78], where not so advanced features were used.

Besides the aforementioned research, there have been some approaches on connecting reliability and evidence with argument strength, but they are not applied in social media

Author	Task	# classes	Algorithm	Metric	Score
Addawood et al. [1]	Evidence Classification	6	SVM	F1-macro	0.83
Dusmanu et al. [36]	Factual vs opinion	2	LR	P	0.81
				R	0.79
				F1	0.80
	Source Identification	NA	str match + h.	P	0.69
				R	0.64
				F1	0.67
Konstantinovskiy et al. [78]	Claim detection	2	LR	P	0.88
				R	0.80
				F1	0.83
		7	LR	P-micro	0.71
				R-micro	0.73
				F1-micro	0.70
				P-macro	0.61
				R-macro	0.44
				F1-macro	0.48

Table 2.6: The results of the supervised ML approaches that have been followed for reliability-related tasks in AM, in text derived from social media. For the task of source identification a rule-based approach is followed. str match = string matching. h = heuristic algorithm

text. For example, a research work that uses argumentativeness as a feature takes place in Cocarascu and Toni [27] for the task of deceptive reviews detection, leading to an improvement of the prediction algorithms. It has to be noted that the exclusive use of argumentative features without topic modelling or the use of additional features cannot surpass the baseline. The work of Park and Cardie [120, 121] is another great example of combining argumentation with evidence classification, suggesting three different categories of justifications in online user comments, but as in Cocarascu and Toni [27], both papers exploit more structured forms of arguments. Similarly, the task of context-dependent claim detection [144, 91] utilizes hundreds of Wikipedia articles, but it has not been applied on social media text.

Another task that is related to both the quality evaluation of the argument and the in-depth analysis on a macro-scale is enthymeme reconstruction. Although the task has not been applied yet in data derived from social media, at least to my knowledge, it has been tested in Web-generated data [136, 138]. Nevertheless, social media text could offer an excellent testing ground for the identification of implicit supporting claims. The task of reasoning comprehension as it has been presented in [60] has the potential to be applied on social media and advance the wider field of AM, but the complexity of the proposed methods raises concerns for the transferability of the task to a new environment and untrained annotators.

2.6 Discussion

The objective of AM is to get a deeper understanding of natural language, while it can be characterised as the natural evolution of opinion mining, but instead of trying to understand what others think, the focus is on understanding why. The analysis of the human reasoning process is the ultimate goal of AM and it is acquired by exploiting the inherent structure of an argument through the identification of its distinctive parts implying the inferential process

that is followed.

The understanding of human reasoning can offer unprecedented capabilities and achieve breakthrough changes in a wide spectrum of applications as information for the decision-making process can be retrieved. The existing work in the field indicates its great potential but we must consider the fact that AM is still in a premature stage and there are steps that are required for realizing human-level reasoning or at least to be able to interpret it on a sufficient level. There are a plethora of tasks under the field of AM, each one answering a different research question and having unique challenges. The research community has focused on the modelling of the argument and the suitable selection of features, whereas the selection of the ML algorithm seems not to play a crucial role in the accomplishment of the different tasks. Additionally, the increase in the number of features increases the total volume of the space transforming the available data into a sparse format.

Recalling the initial research questions that have been expressed at the beginning of the thesis, this chapter covers some significant aspects that are worth researching. More specifically:

- Answers how the field of AM is connected with modern research problems, challenges, and applications through the presentation of related work in the wider field of NLP (research question 1).
- Presents the limitations of the data-driven solutions for the task of argumentation detection (research question 3).
- Partially covers the social implications of argumentation mining expressing its relation with emerging social challenges such as the impact of public debate on social media (research question 5).

Finally, this chapter assists towards the achievement of the thesis' first contribution, it frames the problem of argumentation detection in short text by providing a wider context. It introduces related terms, such as opinion mining and stance detection, and presents the challenges of research when applied to real-life settings. The following chapter delves deeper into argumentation theory, proceeds with the definition of related terms, and proposes the abstract framework for argumentation detection.

Chapter 3

Argumentation theory and modelling

Although AM as a term was first introduced in 2009 by Palau and Moens [119], the act of argumentation and its effects have been studied since the 4th century BC. Since then, many approaches on studying argumentation have been researched and many theories, schemes, and diagrams have been developed. The primary factor that has led to the creation of novel evaluation and visualization techniques for argument representation is the need for simple, but effective ways to break down, analyse and eventually better understand arguments. Argumentation can reach a high level of complexity, thus simpler forms of representation are needed. The process of argument illustration and fragmentation is a fundamental concept in AM, where the arguments are inspected, evaluated and eventually expressed in a binary format, capable of being interpreted by different algorithms. The quantification of argumentation could assist in the deeper understanding of qualitative aspects of the natural language, aiming towards explainable NLP systems increasing the trust towards Artificial Intelligence (AI) achievements. Future modelling schemes should be more agile, and take into consideration the trend of using shorter arguments, providing the necessary flexibility to researchers to implement different versions and compare their performances.

The rest of this chapter is organised as follows. In Section 3.1 significant argumentation theories are presented, followed by the first research studies that realize the connection of argumentation with AI in section 3.2. The proposed conceptual framework, specifically designed for the needs of argumentation in social media, is presented in section 3.3. Section 3.4 illustrates the process for modelling argumentation detection presenting a series of definitions and introducing the AFAD. Finally, section 3.5 initiates a discussion of the findings of this chapter.

3.1 Logical schemes and diagrams

Argumentation diagrams were developed as aiding tools for the analysis of arguments in well-structured documents and they were originally designed to teach the reasoning process without falling into logical fallacies. However, their capabilities as analytical tools for meta-philosophical purposes have also been stressed [141]. Logical diagrams have boosted the field of informal logic, as they offer a tool for analysing and evaluating arguments used in everyday

life in a much more pragmatic environment compared to formal logic.

A detailed review of argumentation diagrams and their connection with other fields such as formal and informal logic, law, and artificial intelligence is presented in the work of Reed et al. [141]. The connection of argument diagrams to modern AM techniques is illustrated in the work of Peldszus and Stede [124], and the introduction of a classification system for argumentation schemes is presented in the review paper of Walton and Macagno [183].

In this section, a synopsis of the five more influential diagrams is presented and they are evaluated based on their suitability for the tasks of AM in noisy text. It has to be stressed that the AM diagrams have not been specifically designed to serve the modern construction of arguments as expressed in social media because the tasks of detection, classification and evaluation of argumentative content in noisy text require more flexible schemes. The need for flexible argumentation schemes is covered in section 3.4 where the proposed framework for argumentation detection in short text is presented.

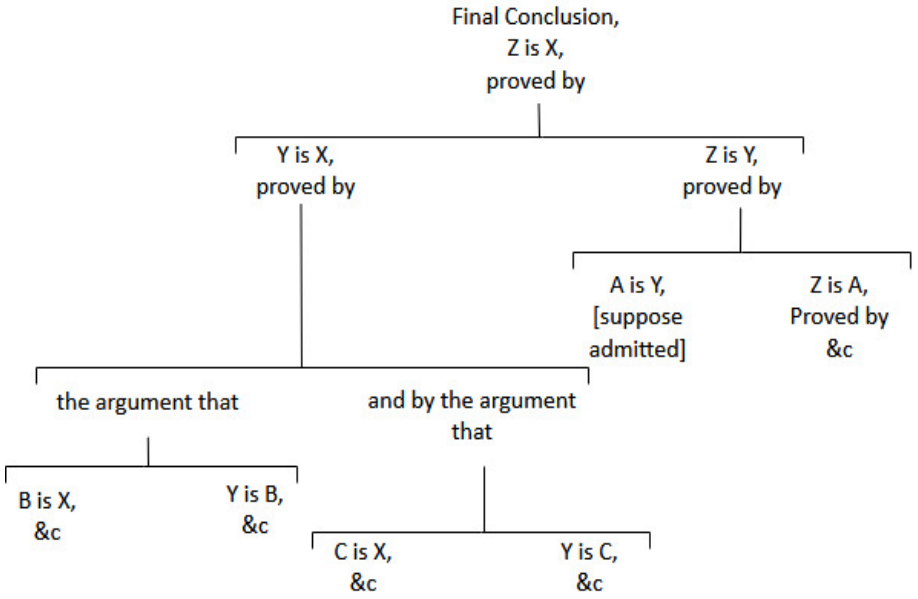

Figure 3.1: Whately's diagram [186] for analysing arguments based on backward reasoning. The final conclusion is represented as the root of the tree and its assertions are represented as leaves and the depth of the tree is proportionate to the complexity of the argument.

One of the first uses of diagrams in argumentation took place in 1857 by Whately [186], opposing the enumeration of technical rules, in an effort of simplifying the teaching method of argumentation in his era. Whately's diagram theory is based on the concept of identifying the concluding assertion and tracing the reasoning backwards, grounding the original assertion and eventually forming a tree with assertions and proofs. In Figure 3.1, the conclusion is represented as the root of the tree, and the assertions are located underneath. In the classical example of Socrates syllogism (Socrates is a man, all men are mortal, therefore Socrates is mortal), the conclusion (Socrates is mortal) would be the root of the tree and the two premises (p1- Socrates is a man, p2 – all men are mortal) would be the two leaves of the tree. The complexity of the reasoning process and the depth of the tree are proportionate and could

lead to a complicated process that requires both well-structured arguments and experienced annotators.

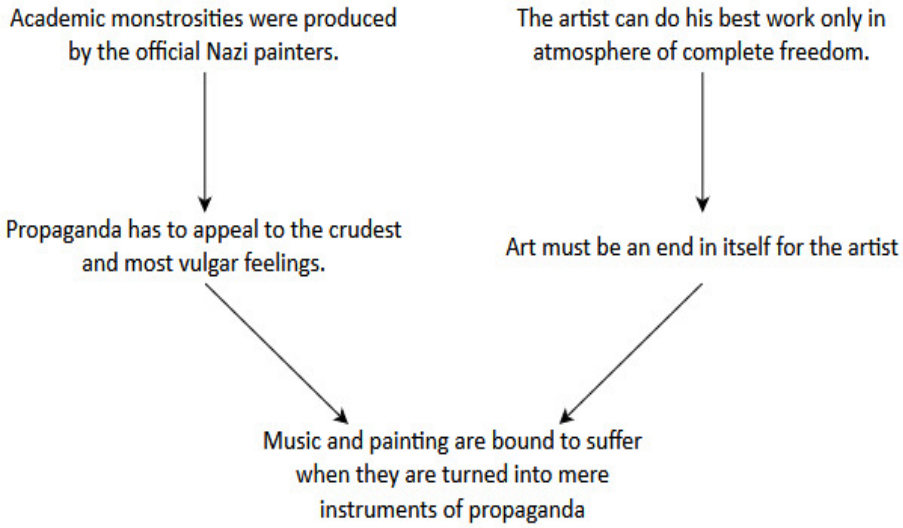

Figure 3.2: Beardsley's convergent argument scheme [8] provided with an example. The serial premises eventually lead (converge) to the final conclusion. It should be noted that the links between the premises are not evaluated.

Another significant work in the field of argumentation is conducted by Beardsley [8] introducing a designing principle that is applied until today; representation of the argument's distinctive statements as linked nodes. As the nodes can be connected with each other in different ways, he created three basic classes 1) convergent, 2) divergent and 3) serial arguments.

Figure 3.2 illustrates a logic flow depicting serial linking between premises, leading to convergence for the final argument. In the example of the convergent argument, different premises eventually contribute towards the establishment of a reliable and robust argument, supported with enough backing. The only flaw in Beardsley's theory is the lack of support between the statements in the nodes. The statements which form the argument are considered flawless and there are no subjects of support, debate or evaluation, hence it cannot be applied in ambiguous, implicit or imperfect arguments.

In 1958, Toulmin suggested one scheme, which is very influential to date, examining the role that different utterances might have in the persuasive perspective of the argument [173]. In Toulmin's model, six functional roles were suggested, namely datum, claim, warrant, backing, qualifiers, and rebuttal.

Figure 3.3 expresses an example of legal nature through Toulmin's model. The warrant in the depicted example (*"A manborn in Bermuda will generally be a British subject"*) adequately supports the initial datum (*"Harry was born in Bermuda"*) as it includes a real strong backing using a legal framework (*"The following statues and other legal provisions"*). The above distinctive components are enhanced through the defy of a possible counter-argument (*"Both his parents were aliens.."*) and eventually lead with certainty expressed through the qualifier (*So, presumably*) to the conclusion (*"Harry is a British subject"*). The novelty of his approach lies in the fact that it requires the assignment of a predefined characterization

for the cognitive connection between the different components of the argument. Through his proposed scheme, Toulmin managed to handle enthymematic relations by defining different aspects of a syllogism and connecting the inference with the warrant. The proposed scheme is widely used in AM and a modification has even been applied in Web-generated data [59], however, the small IAA in some topics indicates the difficulty of applying such a complex model in heterogeneous text.

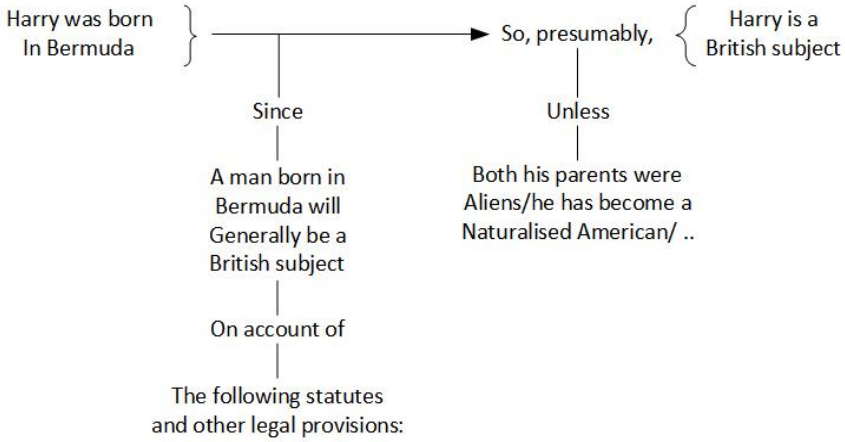

Figure 3.3: Toulmin's proposed scheme provided with an example [173]. Based on its detailed structure, an argument can be assessed through the review of its distinctive components.

Another argumentation theory that still has a strong influence today, as recent research works [81, 125] adopt its scheme in a data-driven approach for AM tasks, is developed by Freeman [43]. Freeman's theory could be characterized as an upgrade of Beardsley's theory, as it uses the scheme of inductive/deductive reasoning and enhances it with the concept of modality, which indicates the strength of induced conclusion by the premises. It could be said that the term modality is an adjustment of Toulmin's qualifier, but with a focus on the evaluation of the argument. Both the concept of reasoning flow and argument strength have the potential to enhance the AM tasks as they offer new unprecedented tasks that have not been tested, as only sentiment flows in AM have been researched [179] until now.

The introduction of a prominent text-organization theory took place in the 1980s by Mann [102], aiming at the organization of the text into different regions. Each region has a central part (nucleus) that is essential for the comprehension of the text, and a number of satellites containing additional information about the nucleus. The nucleus and the satellite are correlated with each other through different relations (circumstance, elaboration, evidence, etc.) which can be changed, manipulated, added or subdivided depending on the topic and the task at hand. The nucleus-satellite distinction is applied recursively until every entity of the discourse is a part of a rhetorical structure theory (RST) relation and eventually a tree-structure hierarchy is created.

Figure 3.4 depicts this tree-structure hierarchy, where the theorem of the perception of apparent motion (initial nucleus) is supported by a set of premises, where each premise sequentially is expressed through the model of nucleus-satellite expressing a specific relation (preparation, condition, means). The application of RST has improved the performance of sentiment polarity classification when enhanced with argumentation [22], but the authors

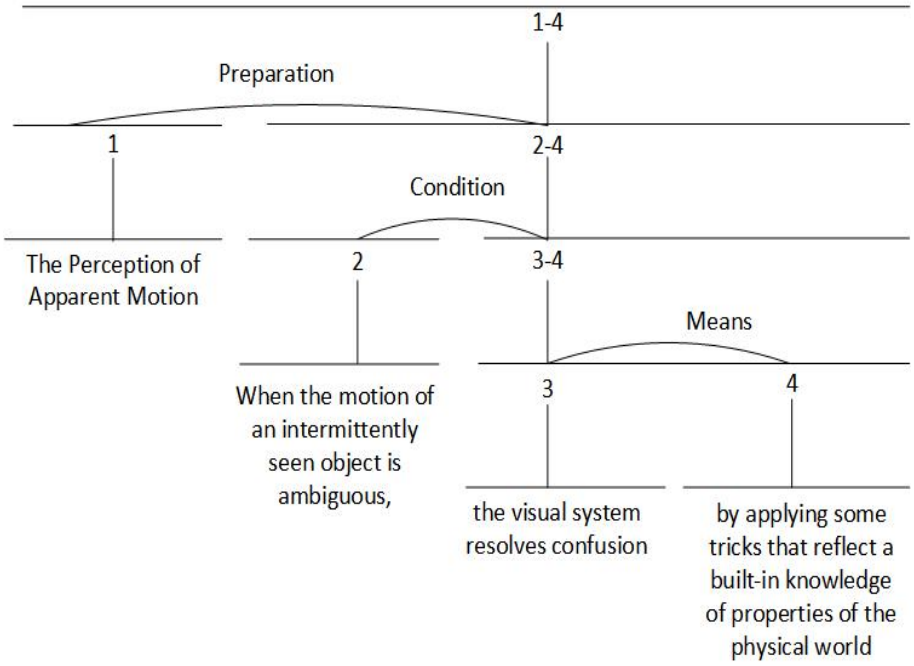

Figure 3.4: Rhetorical Structure Theory scheme provided with an example [102]. The example that is used is the theorem of perception of apparent motion (initial nucleus), which is justified by a set of premises, and each premise is analysed consequently based on the nucleus-satellite model.

state the need for further in-depth evaluation of its impact.

3.2 Early argumentation theory in AI

Argumentation diagrams have boosted the entire field of computational linguistics, but as they are theoretical models, they do not have the capacity to reform the entire argumentation field. In the 1990s, the first research connecting argumentation with the area of AI was conducted. These early attempts on connecting arguments with AI created a new field under the name of computational argumentation, the field which has formed AM to a great extent. Computational argumentation is used extensively in domains where the reasoning process is sophisticated, such as law or medicine, and therefore traditional methods like formal logic and classical decision theory cannot be applied.

One of the first researches studying the relationship between argumentation, cognitive psychology and AI was conducted by Pollock [132] describing the connection of defeasible reasoning in philosophy and nonmonotonic reasoning in AI through the definition of a set of rules. An important step ahead was made by Dung [35], who researched the relation between argumentation and logic programming, focusing on the modelling of the fundamental mechanisms humans use in argumentation and expressing them with a number of rules and definitions capable to be interpreted by AI algorithms.

Similar to Dung's method, Krause et al. [79] developed the logic of argumentation, an

approach for defining reasoning in cases of uncertainty. By defining rules, definitions and propositions they have managed to create a theoretical system capable of estimating the strength of an argument based on the distinctive axioms used for its construction. The degree of justification was also researched by Pollock [133]; focusing on the different degrees of justification the distinctive components of an argument can provide when they are *"summed"*.

Another work that indicates the relationship between argumentation and logic programming was held by Parsons and Jennings [122], where a framework was developed capable of interpreting a broad range of negotiation in a multi-agent system. The distinctive arguments of the agents are considered as logical steps towards an acceptable compromise, not necessarily towards the optimum proposal. A more detailed review on the automated negotiation is held in Jennings et al. [70] presenting a more generic framework and three different approaches (game-theoretic models, heuristics, argumentation-based) for the negotiation process. The development of a dialectical argumentation scheme was also introduced, where the authors implemented a fully-functional agent capable of maintaining the natural flow of the debate by arguing on a topic and at the same time responding to obligations related to the discourse [52].

Author(s)	**Link Relation**	**Complexity Level**	**Application in noisy data**	**Application in AM**
Whately [186]	No	High	Low	Medium
Beardsley [8]	No	Low	Medium	High
Toulmin [173]	Yes	High	Medium-Low	High
Freeman [43]	Yes	Medium	Medium	Medium
Mann [102]	Yes	Open	High	High
Pollock [133]	Yes	High	Low	Medium
Dung [35]	Yes	High	Low	Medium
Krause et al. [79]	Yes	High	Low	Medium
Parsons and Jennings [122]	Yes	High	Low	Medium
Grasso et al. [52]	Yes	High	Low	Medium

Table 3.1: A synopsis of logical schemes and computational theories based on the degree they can meet the needs in a modern NLP environment. Link relation - if the connection distinctive components are explicitly evaluated, Complexity level - is assessed based on the number of components each theory includes.

Although many of the aforementioned frameworks are designed to cover argumentation in its full generality, any system with pre-defined rules is not suggested for use in the open and constantly changing environment of social media, as it does not have the ability to learn and adapt.

Table 3.1 presents a synopsis of the theories and schemes that have been examined in this section. The theories are evaluated based on a number of criteria, focusing on their suitability on AM in social media. The table could also be used as a point of reference for tasks in the wider area of NLP. The first criterion of a theory's evaluation is the explicit expression of any kind of relations between the distinctive components of the argument, a property that can facilitate advanced NLP tasks such as relations identification, evidence

detection, and facts recognition. In the theories that are based on logic programming, the relations are expressed as rules, whereas in Toulmin's theory they are expressed through the qualifier and in Beardsley's as modality. The second criterion in the comparison table is the level of complexity, which is determined based on the number of components and relations each theory involves. Both the in-depth analysis and the construction of the reverse tree in Whately's theory, and the definition of six different components in Toulmin's diagram present high levels of complexity, as expected. Once again logic programming theories can be grouped together presenting high complexity as they define multiple rules and axioms in order to cover a wide range of cases. A special note should be given in RST, as it is a flexible open scheme, where components and relations can be defined and modified based on each case study. The first assessment of the theories/schemes takes place based on their suitability to be applied in NLP tasks in a noisy environment. The complexity level of a theory is inversely proportional to its suitability when applied to a noisy environment as complicated reasoning processes cannot be deployed in text lacking grammatical and structural rules. Finally, the last column of the table illustrates the applicability level of the theories for AM tasks independently from the source of the content, where three schemes stand out (Beardsley, Toulmin, RST), each one for different reasons. Beardsley's theory is straightforward enough to be applied easily to a wide range of goals, Toulmin's detailed scheme is the option for in-depth analysis and RST can be easily modified in order to cover the needs and the requirements of each case study.

Computational argumentation contributed significantly to the creation of the AM field, but the strict rules that used to be posed seem outdated. Although the foundations of logic programming are based on knowledge- and logic-based solutions, the recent advances in ML have been used extensively in data-driven approaches [161, 2]. However, they seem to have reached the upper limit of their capabilities. Rule-based machine learning could be considered as a data-driven approach of inductive logic programming that combines background knowledge and the ability to learn based on human-readable theories. Another approach combining a data-driven solution in unison with argumentation structure took place in Carstens and Toni [22] where the probabilistic classifiers have been improved with the incorporation of an argument database.

3.3 Proposed conceptual framework for AM in social media

The two most dominant characteristics of text derived from social media are the short length and the lack of defined norms. Considering the fact that the typical length of a tweet is less than 50 characters, the definition of argument boundaries in many cases is not feasible. An argumentative tweet does not have the capacity to contain information unrelated to the major claim. On the other hand, the already unstructured nature of text data in combination with the massive use of jargon and emoticons establish an environment where it is really challenging to apply any lexical rules. They have to be specific for each case study or loosely defined in order to include different cases. Both assumptions would probably lead to a lack of transferable knowledge, a crucial objective for almost every proposed methodology.

A possible criticism for the lack of the boundaries-definition task as defined in Lippi and Torroni [96] is the increase of the upper limit in Twitter from 140 characters to 280. Additionally, there are various social media platforms including forums such as createdebate.com, where the norm indicates well-structured arguments and lengths significantly longer than 50

characters. However, the great dominance of Twitter in socio-political issues, which is also boosted by the online presence of political leaders, combined with the shrinking of general-purpose forums, have formed an environment in which the use of social media data seems to be the only source of Web-generated data in the near future. The increased use of social media in the wider area of AM as a source of information highlights the need for the definition of a new scheme devoted to the specialized procedures in the field of AM in social media.

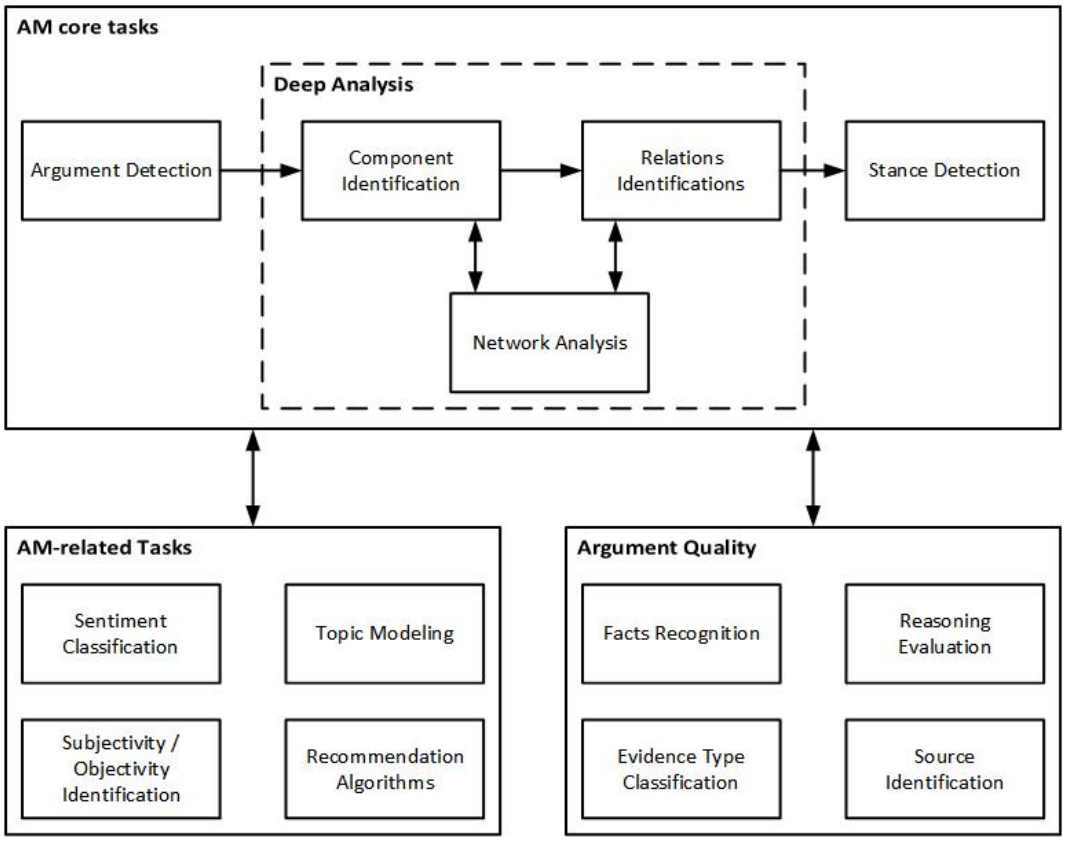

Figure 3.5: The Proposed Conceptual Architecture for AM

Figure 3.5 depicts the proposed conceptual architecture. It contains three main components and each one includes distinct tasks that could form a second focused pipeline, but could also exist independently from the other tasks. The first component contains the core tasks of an AM procedure, the tasks that make the heart of the pipeline and can be applied to different text data, from official political speeches to tweets and comments in products or services reviews. The other two components include tasks that are involved somehow in the process of evaluating the reliability of the text. The additional components should not be underestimated and should be considered of equal importance compared to the AM core task. The interest of the scientific community is constantly increasing in tasks such as fact-checking, evidence and fake news detection, even though the connection between AM and reliability-related tasks is not yet very solid.

The first step of the first component in the proposed conceptual architecture is the task of argumentation detection, where the identification of a sentence as argumentative or not

takes place. A significant amount of work intentionally ignores the step of argument detection [160, 83], as argumentative rhetoric is a prerequisite for the evaluation of persuasive essays or political debates. When collecting heterogeneous data from social media the detection of argumentative text is unavoidable, as not all users intend to persuade for or against a discussed topic, but simply express a reflection, a feeling or a question. The task of argumentation detection is considered an essential step for the AM pipeline in social media, as the following steps of the pipeline are not possible to be completed if it is missing.

A great amount of research work focuses on the identification of the components that construct the argument, especially the early attempts of argumentation (see section 3.1). These attempts present as the ultimate goal the successful analysis of the argument, emphasizing the reasoning concept behind the structure of the argument. The concept of in-depth analysis of an argument through the identification of the argument's components and the discovery of the underlying relations between them has been adopted and developed in the field of AM in social media, adjusting to the new environment including the relations and the interaction between the entities of the network. The two main tasks that fall under this category are the relation-based AM [22] and the enthymeme reconstruction [136]. Both tasks provide useful insight for the structure of the argument, but instead of trying to evaluate its impact, they focus on the efficient use of small datasets [138, 27] or they are designed to aid the task of stance detection [136].

Earlier in the thesis, the task of opinion mining has been characterised as the predecessor of AM, mostly because the terms stance detection and opinion mining can be used interchangeably. Stance detection is the final step of the proposed framework and it is a task that is heavily dependent on the previous steps of the pipeline, as the nature of social media data demands a significant pre-process procedure. This procedure in AM pipeline is expressed through the detection of argumentative tokens and the identification of the relationship between the components of the argument (explicit or implicit).

The constant generation of data in social media has raised significant concerns about the quality of information that is shared and read on social media. The connection between argument in social media and reliability has not yet been discovered in-depth, compared with the amount of work that is dedicated in the areas of rumour and fake news detection. However, tasks such as evidence type classification, source identification and facts recognition have emerged in the area of AM, raising the awareness of the connection between the expressed argument and its reliability.

In the proposed conceptual architecture for AM in social media, a unique component of the pipeline in reliability-related tasks has been devoted, in order to stress its importance and the room for the development that exists in the area. Reliability-related tasks are not considered core tasks of the AM pipeline, but they can enhance the procedure, especially when applied to arguments derived from social media, as the backing of the claims is many times inaccurate or it is based on rumours or hoaxes. Other tasks that can be assisted by the progress in the field of AM and integrate parts of the proposed AM core tasks are sentiment classification [111], subjectivity/objectivity identification, topic modelling [181], and recommendation algorithms [75].

Different tasks can be easily executed and combined through the components described in our proposed architecture, as it is both detailed and easily modifiable. The example provided in 3.1 concerning the Apple/FBI encryption debate (*RT @ItIsAMovement "Without*

strong encryption, you will be spied on systematically by lots of people" - Whitfield Diffie), can be assessed for its argumentativeness nature, the relations that were expressed through the retweet and the mention, its stance towards the discussed debate, its completeness and integrity, while other NLP-related tasks can also be enhanced from the findings of the above tasks. The intention of the proposed framework is to be regarded and used as a means for enhancing various NLP tasks with the use of argumentativeness features.

3.4 Argumentation modelling

Generally, modelling tasks in social media text, such as fake news detection and rumour diffusion, does not have a theoretical framework that provides solid knowledge for the different tasks, however, recent research on the field indicates the potential of modelling. There has been recent research providing detailed modelling strategies on rumour diffusion defining and implementing a multi-feature diffusion model [56], while a different study proposes novel designs mitigating adverse effects caused by the sparse target data samples [188]. The internal structure of false narratives is also an interesting perspective providing insights on the complex link between different independent factors [66, 4], and the modelling and design of a truth campaign offer a viable alternative to blocking nodes [62]. Nevertheless, there has not been any effort, at least to my knowledge, to provide specific definitions and propose an argumentation framework on argumentation in short text.

This section presents the challenges of argumentation in short text in subsection 3.4.1. Then, it proceeds to present the related definitions for the quantification of argumentation and proposes a framework for argumentation detection in subsection 3.4.2.

3.4.1 Argumentation in short text

As in any multidisciplinary field, argumentation is open to different definitions and interpretations depending on the scope of each research study. In this thesis, an argument is defined as a series of statements expressed in natural language, called premises, intended to support, and eventually determine the effectiveness of another statement, the conclusion. The above definition, although solid and complete, is elaborated and modelled in this section to fit the requirements of logic, computational argumentation, and, eventually AM.

An argument's computational model can automatically assign strength values to a statement by evaluating various aspects of argumentation such as persuasion, cohesion, and stance detection. The success of the abstract argumentation framework [35] is merely due to its ability to get extended and modified to assign numerical values to a statement and eventually offer a natural link to statistical methods and tools. This ability of a computational model to quantify a concept that has qualitative substance is of major importance because it can provide a crucial incentive to the development of human-level reasoning machines capable of interpreting human argumentation. Focusing on the field of AM in social media, an explicit definition of argumentation (or argument) has not been given, since different scholars adopt different views and avoid providing a strict definition of argument.

Different norms and deduction processes are followed in social media text compared to the argumentation rules that are observed in formal discussions, such as political debates and legal affairs. Arguments in informal discourse rarely follow the logical structure of an argument

where claims are supported by facts, warrants and qualifiers. On the contrary, arguments on social media are often implicit and without a solid logical structure, thus more agile approaches have to be followed to detect argumentation in a statement and further analyse it. For example, an argument expressed as a tweet is typically a one-sentence argument expressing a stance supported by a premise that either tries to provide a valid reason to justify the expressed stance or it supports it with facts, which are often expressed as external resources.

Having in mind those special characteristics of argumentation in short text, a new agile definition for argument must be provided, covering the needs of this emerging area. In this thesis, before elaborating on the essence of argumentation from a computational perspective, a short mathematical expression is provided, forming a definition strict enough to exclude ambiguities, but at the same time fairly agile to be applied in noisy text.

Definition 1. ***Existence of argument*** is expressed as a quadruple in: $\langle s_{\text{ijt}}, c_{it}, r_{jt}, t\rangle$ where s_{ijt} is the initial statement or set of statements that is supported by c_{it} claims (or premises) that are used to support or oppose an idea (or suggestion) for the topic t that is questionable or open to doubt using a rationale r_{jt}.

The definition 1 covers a wide range of argumentation variations in social media discourse as 1) it narrows down the area of interest of a selected topic t, 2) identifies the stance (positive, negative) of the claimant through the claim c, and 3) proceeds -to some extent- to the examination of some urgent topics in the NLP area in social media text such as reason acceptability, facts recognition, and rumour detection through the inclusion of reasoning r in the definition. The subscripts ijt depict the direct relations between the components. For example, the claim c is a structural component of the statement s (declared through i) for a specific topic t, but there is not an explicit relation with a reason r.

Considering the noisy nature of text on social media and the challenges involved, AM seeks to create a novel approach to the problem, attempting to assign quantitative features to qualitative characteristics. Adjusting the definition 1 in the wider context of NLP on social media, a tweet is equivalent to a statement, the presence of hashtags poses the limits for the topic selection, and the identification of claims is accomplished through the task of stance detection, and the reasoning detection is a combination of tasks such as source identification, rumour diffusion and reliability evaluation.

The noisy nature of text in social media poses a series of challenges, hence argumentation modelling schemes should seek novel approaches to the problem to map qualitative characteristics to quantitative features. The proposed modelling schemes should be more agile, providing the necessary flexibility to researchers to implement different versions and compare their performances. Furthermore, by formalizing the definition of argumentation in short text, the theoretical foundations for different development strategies are provided, assisting in the development of experiments that can produce comparable results.

3.4.2 Argument quantification

A statement, a tweet in the case of social media text, is considered argumentative when it provides reasons for or against the discussed topic. After studying a series of short statements on multiple topics, it is concluded that two different rationales fall under this definition: 1) (try to) persuade towards a specific stance, and 2) provide evidence (news media, blog,

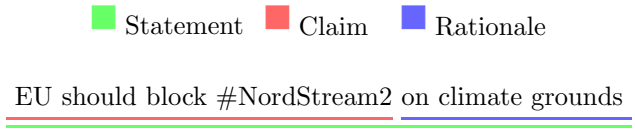

Figure 3.6: An example illustrating the existence of argumentation as given in the definition 1 on short text.

expert opinion) that supports a stance towards the discussed topic. An example of the given definition is illustrated in Figure 3.6. The tweet is a statement for the construction of the gas pipeline under the name *#Nordstream2* (topic), expressing a negative stance through a claim (*EU should block*), which is justified by a reason (*on climate grounds*). The argument, although short, is solid and on point since it is compliant with the definition's requirements for its existence.

The reliability of the reasoning is not examined, because the integrity and the soundness of the argument are not in the scope of this thesis. In the environment of social media, where different topics are discussed, trends appear, and hashtags are created in the blink of an eye, it is important to narrow down the scope by setting limits. In this constantly changing environment, the limits are defined from the topic each statement addresses, thus it is important to define the concept of the topic.

Definition 2. ***Topic*** A topic t determines the context of the discourse under examination and defines the dialectic limits that can be applied in a logical acceptable statement. The finite potential topics formulate the set T, and therefore $t \in T$.

The limits of each t can be differentiated depending on the needs of each case study. A prerequisite for argumentation is the existence of claim(s) in the block of text under examination.

Definition 3. ***Claim*** A claim c is an assertion containing or implying a stance towards a topic t. The finite potential claims formulate the set C, and therefore $c \in C$.

Through the presence of a claim, the stance of the claimant towards the topic t can be extracted. The stance has already been defined [100] and although similar to the Definition 1, it is the product of argument analysis and not an integral part of it. The goal of the claimants is to persuade their audience, supporting their original claim(s) through a reasoning process that completes the argument. Even in very short text, the existence of justification is still very important, even if it is incomplete or weak.

Definition 4. ***Reason*** A reason r is a justification that supports or tries to support a specific claim c towards a topic t. The finite potential reasons formulate the set R, and therefore $r \in R$.

The co-existence of a concrete claim followed by a reason does not necessarily create a legitimate argument, because the provided reason may not sufficiently support the original claim. Therefore, a mapping function assessing the appropriate match of the two objects needs to be defined, demonstrating that every reason r can sufficiently support a specific finite number of claims c.

Definition 5. ***Mapping reason to claim*** A reason is valid when there is a function $f_{rsn}(r_i, t) : R \rightarrow C$ valuing the validity of a generic reason r from the predefined set of reasons $r \in R$ to support a a claim c from the predefined set of claims $c \in C$. Then, the valid reasons for a claim are expressed as a set $Rc = (r_1, t), ..., (r_n, t)$ (for $n \geq 0$) and a mapping function v on Rc is defined as:

$v(r, c) = g((r_1, t), ..., (r_n, t))$,

where $g : R \times T \rightarrow C$ is a function aggregating and assessing the impact of the existing distinctive valid reasons capable of supporting to support assertive claims.

Following the definitions of an argument's components, the definition of the argument itself follows using computational logic. Assembling the distinctive components of an argument, the argument's definition is provided capable of capturing the presence of argumentative discourse in a statement and of being applied to noisy text.

Theorem 1. An argument a is a statement that supports or tries to support a specific claim c where $c \in$ C towards a topic t, supported by a reason or set of reasons $R_{ct} \in R$ with $R_{ct} \geq 1$ and it is expressed as a set of tuples $\{(c, r_j), (c, r_{j+l}), ..., (c, r_{m-l})\}$ where $0 \leq l < m - j$, $\forall j$ where $1 \leq j < m$ iff $\exists\ c \in C \bigwedge \exists\ r \in R_{ct} : (c, r) \in A$, where A is a finite set of the potential arguments.

The above theorem introduces the prerequisites for the creation of an argument adopting a computational approach, but it does not evaluate its validity which is assessed through a function that maps the suitability of the claims to the argument as an entity.

Definition 6. ***Mapping claim to argument*** A claim is asserted when there is a function $f_{clm}(c_i, t) : C \rightarrow A$ valuing the applicability of a generic $c \in C$ to create an $a \in A$. Then, the asserted claims for an argument $a \in A$ are expressed as a set $Ca = (c_1, t), ..., (c_l, t)$ (for $l \geq 0$) and a mapping function u on Ca is defined as:

$u(c, a) = h((c_1, t), ..., (c_j, t))$,

where $h : c \times T \rightarrow A$ is a function aggregating and assessing the impact of the existing statements. The statements are considered arguments because they consist of valid reasons and assertive claims.

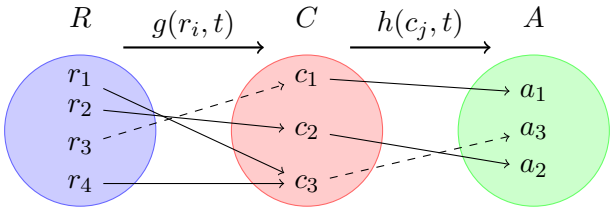

Figure 3.7: A graphical illustration of mapping reasons to claims and eventually to arguments.

Figure 3.7 illustrates the mapping process from a set of reasons to a set of claims through a function $g(r, t)$ and from a set of claims to a set of arguments through a function $h(c, t)$. The two functions have as common the input parameters for the same topic t which defines the limits of the sets and both $g(r, t)$ and $h(c, t)$ present $n : 1$ and $1 : 1$ relation from input parameters to output, respectively. The relation $n : 1$ indicates that multiple reasons could

support the same claim, while the 1 : 1 relation expresses that an argument can have only one claim.

The solid lines between the sets indicate a direct link between them, whereas the dotted lines represent an implicit correlation without any evident connection. Neither the argument definition nor the mapping functions examine the validity of the argument, but they are limited to its detection and the identification of their distinctive components. Therefore, a definition determining the validity of an argument is also required.

Lemma 1. An argument for a topic $t \in T$ is a tuple of the form $\{c, R_{ct}\}$ where $c \in C$ and $R_{ct} \in R$ is a set of reasons supporting the claim c while $R_{ct} \geq 1$. An argument as a whole is valid if at least one element of the R_{ct} is valid and implies a claim $c \in C$, and c creates an argument a, and it is expressed as $a : (r, r \rightarrow c, c \rightarrow a, \therefore a)$.

Proof. The expression is proved from the law of detachment. For the validity of the claim we have $(r \rightarrow c, r, \therefore c)$, and similarly the validity of the argument is proven $(c \rightarrow a, c, \therefore a)$. □

In other words, an argument is valid when it consists of a reason that is based on sensible reasoning and assertive claims. Apart from validity, another important aspect of argumentation is the soundness of an argument; an argument is sound when it is valid, and all of its premises are accurate and truthful. Symbolic logic can be used to check the validity of an argument, but it cannot be used to examine its soundness. The detection of a sound argument is in the scope of research of evidence, including tasks such as source identification, facts recognition and evidence classification.

Finally, having recognized the distinctive components of an argument and provided definitions for each one of them, the Abstract Framework for Argumentation Detection (AFAD) is introduced. It is a conceptual model that can be applied to (very) short text and detect the presence of argumentative text. The AFAD is the first framework -to the best of my knowledge- that focuses on the detection of argumentation and not its evaluation, while its structure allows its utilization on short and noisy text. In contrast to previous argumentation frameworks, AFAD offers a great level of flexibility due to its ability to be adjusted to the needs of every experiment, allowing researchers to define the mapping functions.

Definition 7. ***Abstract Framework for Argumentation Detection (AFAD)*** is a 4-tuple $\langle C, R, T, V \rangle$, where C is a finite set of claims that are supported from a finite set of reasons (justifications) R. The parameter T indicates the topic for which the C and the R are expressed; a necessary parameter in order to narrow down the potential tuples of $\langle\ C,\ R\ \rangle$, and $V : A \rightarrow C \times R \times T$ is a function mapping valid arguments to their building components.

The absence of transition words and the overlapping between claims and conclusions in short text create an environment where original ideas and schemes of AM can be applied. Therefore, if there is the need to juxtapose novel approaches to established argumentation schemes (e.g. Toulmin's model), fillers should be added to reconstruct a complete argument.

3.5 Discussion

The field of AM has the potential to stimulate a series of applications, where the evaluation and the classification of reasoning are essential, especially in Web content, where both

reliability and reasoning validity of a user holding a position are questionable. Tasks such as troll detection, knowledge retrieval, and information validation could significantly benefit from the progress in AM. Cutting-edge techniques in human-computer interaction, opinion mining, and recommendation systems could adopt parts of an AM system and enhance their performance. The successful interpretation, evaluation, and taxonomy of arguments will eventually lead to human-level reasoning machines, which can understand, evaluate, and eventually create knowledge. Different approaches have been tested for modelling arguments from diagrams depictions to modifications of well-established theoretical models and no model excels in comparison to others, creating the need for establishing a flexible framework able to capture the needs of the different tasks that appear in AM problems.

This chapter tries to shed light on the argumentation field, to provide a clear view of the wider area with a focus on automatic AM in social media text. Different models were presented and analysed to their core tasks. Inspired by the individual AM sub-tasks, a complete conceptual architecture for AM is presented. The proposed architecture can be easily adapted and modified depending on the goals of every research work. Furthermore, by formalizing the definition of argumentation in short text, the theoretical foundations for different development strategies are provided. A formal definition of argumentation is introduced providing the necessary theoretical background for different implementations. The findings of this chapter can be used as a point of reference for future studies by exploring different implementation methods in the proposed modelling scheme.

Recalling the initial research questions that have been expressed at the beginning of the thesis, this chapter covers some significant aspects that are worth researching. More specifically:

- Presents the technical exploration of the argumentation modelling field illustrating how argumentation has evolved and expressing its relation with modern research challenges (research question 1).
- Illustrates the changes of argumentation schemes throughout the years and highlights the changes social media has brought in language and argumentation (research question 2).
- Stresses the impact of social media in argumentation and public debate, partially covering the social implications of argumentation detection (research question 5).

Finally, this chapter assists in the achievement of the thesis' first contribution by framing the problem of argumentation detection in short text. It proposes an architecture for future AM pipelines and a theoretical framework for argumentation detection in short text that can be deployed in real-life settings. Additionally, it assists towards the second major contribution of the thesis by exploring argumentation techniques in short text. The following chapter presents related datasets in the task of argumentation detection and illustrates the creation of the social media dataset.

Chapter 4

Annotation process and dataset creation

The previous chapter reviewed existing argumentation schemes, presented the challenges of argumentation in short text, provided a series of definitions, and proposed a framework of argumentation detection. The construction of a dataset on argumentation detection in short text includes a series of challenges from the selection of annotators to the final format of the dataset. This chapter focuses on providing details on the annotation process that is followed, how it is connected with the theoretical argumentation framework, and discusses the challenges that appear in the annotation process of shorter chunks of text and the possibility of its application on a different domain to provide the context for the annotation objectives and the related AM tasks, such as evidence recognition and stance detection.

The rest of this chapter is organised as follows. Subsection 4.1 provides an overview of related work on the construction of datasets in the field of AM, subsection 4.2 describes the annotation process that is followed for the construction of an argumentation detection dataset on the public debate for the construction of the Nord Stream 2 in Twitter, and its connection with the theoretical argumentation modelling. Finally, section 4.3 initiates a discussion of the challenges and findings of this chapter.

4.1 Annotation process and available datasets

The field of AM aims at extracting valuable information from natural language. This thesis focuses on argumentation detection in short statements, emphasizing the challenges that occur when argumentation detection techniques are applied in real-life settings. Towards the analysis of a short statement, the first step is to set the level of background knowledge that a reader/annotator should have for the discussed topic in order to ensure the quality of the annotation. To facilitate the process of the annotation task some general rules are defined to provide a better understanding of the nature of the text. The rules are defined depending on the context, the aim, and the objectives of every research project. However, some common guidelines can be easily agreed upon. For example, when the user cites the opinion of an expert, it is an appeal to authority (argumentum ab auctoritate), thus the tweet is very likely to be argumentative.

On the other hand, if the tweet is the title of an article and simply shares a link without

any comments, the tweet can be characterized as non-argumentative. An example of a non-argumentative tweet is provided in the work of Dusmanu et al. [36] regarding Brexit: *72% of people who identified as "English" supported #Brexit (while no majority among those identifying as "British") https://t.co/MuUXqncUBe*, where the statement does not include any explicit or implicit argument for or against the Brexit debate. Instead, it simply presents a fact from a survey.

The rules that have been described should be considered more as guidelines since there are many cases where a tweet cannot easily fit in any of the proposed categories. In many cases, a Twitter user takes into consideration more information than simply the text contained in the tweet, such as the beliefs or the status of the user who tweeted. Therefore, defining a specific level of knowledge of the annotators and some ground rules is crucial for every NLP experiment.

Other important factors that affect the annotation process are the quality of the dataset, the task at hand (argumentation detection, source evaluation, etc.), the size of the dataset, the debate topic, and the number of annotators that have been involved. The above characteristics define the quality of the annotated dataset which is evaluated using the metric of the Inter Annotator Agreement (IAA). The IAA is a measure of assessing how well the annotators can conclude into the same annotation decision for the given task, revealing how clear the annotation guidelines are, how uniformly the annotators respond, and, eventually, if the annotation task is reproducible. It is a crucial metric for the quality of the dataset and the continuation of the experiment, hence it is suggested to be reported in every experiment. The two most popular metrics are Cohen's kappa [28], and Krippendorff's alpha [80]. Both of them are considered more reliable than simple percentage agreement calculation because they take into account the possibility of the agreement occurring by chance.

Argumentation mining compared to other NLP tasks is still in its early steps, hence there is a lack of annotated datasets, especially in short text. The shortage of datasets hinders the extraction of best practices and the definition of annotation guidelines creating an ambiguity in the annotation process. Focusing on social media platforms, users often express arguments without solid or even missing justifications, a rhetorical practice that has not been met in professional writing and which was the focus of early argumentation schemes. The development of users' networks with common characteristics that discuss specific issues favours the use of domain-dependent idioms and self-references, creating a hostile environment for the annotation process. In a constantly changing environment, where the nature of argumentation is strongly correlated with the topic and the relationships that are created within the network, forming rigid annotation rules that can be applied on different contexts is far from a trivial task.

Four basic annotation guidelines for argumentation detection in Twitter data are provided in the DART dataset [17] which can be further elaborated and used in different contexts. The first one is that an opinionated tweet is considered an argument, without seeking a reason to determine its validity. According to the second annotation direction, claims that are expressed as questions are considered argumentative, regardless of their objective and goal. The third guideline recognises as argument any tweet containing factual information that could be characterised as a premise or conclusion. Finally, when a tweet contains factual information and it can be understood without requiring external resources, then, it is considered as self-contained, and thus argumentative.

The available datasets on argumentation mining in short text are limited because it is an expensive process and scholars offer them after the completion of their research. However, the annotation process, quantitative features, and some qualitative ones are provided. It is important to provide the available information for the creation of the datasets, study the input data, identify any implicit bias, and, eventually, increase the trust in the outcome of the AI methods.

Authors	**Size**	**Topic**	**IAA**
Addawood et al. [1]	3000 tweets	Apple/FBI encryption debate	0.67 Ck
Bosc et al. [18]	4000 tweets	5 debate topics	0.81 Ka
Dusmanu et a. [36]	1187 tweets	Grexit / Brexit	0.77 Ck

Table 4.1: A comparison of IAA for the task of argumentation detection

Table 4.1 provides a review of the characteristics of the datasets that have been used in recent research papers that undertake the task of argumentation detection. The first column cites the related work, the second column has the size of the dataset, the third one is the topic of each dataset, and the last column has the score of the IAA. The task of argumentation detection, as a necessary preliminary step, takes place in the work of Addawood and Bashir [1] and also in Dusmanu et al. [36], where the proposed pipelines result in the recognition of evidence type and in the source identification, respectively. In both works, the detection of argumentative tweets is necessary, as the non-argumentative tweets account for 42.3% and 29.3% of the constructed datasets, respectively. Both papers adopt a supervised approach where a manual annotation is required, succeeding a substantial IAA in terms of Cohen's kappa equal to 0.67 and 0.77, respectively. A similar architecture was also adopted in Bosc et al. [18] ending up in the construction of argumentation graphs. The DART dataset [17] was used as input of the pipeline, containing 2702 argumentative tweets and 1181 non-argumentative tweets. The IAA was measured in terms of Krippendorff's alpha resulting in the satisfactory alpha=0.81.

The aforementioned studies perform the annotation process for the task of argumentation detection, and they proceed to further analyse the input text applying different tasks. Quantitative aspects of the annotation process in AM-related tasks, such as evidence categorization and relations identifications, are presented.

Authors	**Task**	**IAA**	**#Classes**
Addawood and Bashir [1]	Evidence categorization	0.79 Ck	6
Bosc et al. [18]	Relations Identification	0.67 Ka	3
Dusmanu et al. [36]	Factual vs opinion	0.73 Ck	2

Table 4.2: A comparison of IAA for the task similar close to argument detection

Table 4.2 presents the IAA in related work for tasks that are semantically close to argumentation detection. It is important to provide related AM tasks to better understand the challenges that appear in the annotation process. Based on the findings, the IAA depends heavily on the task at hand, as it is revealed in Addawood et al. [2], and Addawood and

Bashir [1]. On those two research papers, the same annotators in the same dataset, present higher IAA for the tasks of evidence/topic classification in comparison to the tasks of argument/stance detection, although the latter offers more possible classes than the former. The last point that should be stressed is the use of different metrics since different studies choose different annotation strategies.

Expanding the boundaries for the initial sources beyond social media data, there is significant work that is worth mentioning. For example, the transcripts from the two-hour debate aired by Sky News on April 2, 2015, having statements from David Cameron, Nick Clegg, and Ed Miliband [95]. The text from the political debate offers a different perspective allowing us to test the proposed methodologies on new domains and topics. Providing more details on the dataset, the transcripts are framed representing self-contained sentences having eventually 386 statements, 122 for David Cameron, 104 for Nick Clegg, and 160 for Ed Miliband. Regarding the IAA, Cohen's kappa has reached 0.52 for Cameron, 0.57 for Clegg, and 0.47 for Miliband, while the combined kappa was equal to 0.53, reaching an overall "fair to good" agreement.

Related tasks in the wider area of AM can present higher variance depending on both the task that takes place and the nature of the dataset. For example, for the task of matching argument to specific key points, a moderate agreement (0.5 Cohen's kappa) is achieved in Bar et al. [7], whereas for the tasks of reasoning revision on argumentative essays [3] and identification of argumentative components in medical abstracts [106] moderate agreement is achieved, with 0.75 Cohen's kappa and 0.72 Fleiss' kappa respectively.

In order to provide a complete overview of the wider area of annotation, some alternative annotation approaches follow. A semi-supervised approach was followed in the papers presented at the CLEF 2018 conference for the task of multilingual cultural mining and retrieval. In the work of Deturck et al. [32], the hypothesis is that a text is argumentative when it is structured to effectively combine arguments and opinions, thus argumentation is measured through structuration. A similar approach was followed in Sendi and Latiri [154], where argumentative tweets are defined as the sum of three separate tasks: information retrieval, topic modelling and sentiment score. The work from Dufour et al. [34] follows a distance supervision approach through the detection of five pre-defined features: emotion words, emoticons, particular punctuation signs, personal pronouns and hashtags.

Overall, there is a scarcity of annotated datasets in AM and most importantly there is a difficulty in their re-use, as they are often annotated for very specific tasks. The recent review paper of Cabrio and Villata [21] provides a complete synopsis of the available datasets in the area of AM, but without focusing on short text, Web-generated or social media text. There is an evident need for publicly available datasets in the task of argumentation detection in order to engage the research community without the need for from-scratch creation and annotation of new datasets. The dataset that is created in this thesis is available upon request.

4.2 Creation of the Nord Stream 2 dataset

The classification of a sentence, or a series of sentences, as argumentative or non-argumentative is a crucial step towards AM in real-life settings. Argumentation detection is essentially a preliminary binary classification that would enable the subsequent in-depth analysis of the argument, such as persuasiveness detection, relations identification between the components

of the argument or automatic evaluation of the argument. Argumentation detection is often ignored when debate scenarios are assessed, but it is of great importance when it addresses real-life scenarios such as public debates on social media where arguments, facts, opinions, and fake news coexist. The outcome of this process may be useful in different contexts, like legal reasoning or public legislation, and most importantly in controversial topics.

Focusing on data from social media, they are characterized as a special category of Web-generated data and this becomes clear if an attempt is made in applying the Toulmin scheme to an ordinary tweet. Additionally, due to the crucial role of social media as means of communication, the use of natural language has undergone significant changes, altering the norms and rules. In short, chunks of text in traditional argumentation schemes, such as Toulmin or RST [102] cannot be applied, and as a result, the prerequisite for a statement to be characterized as argumentative is the existence of a rationale supporting a claim. The validity of the rationale lies in the sphere of source reliability and fake news detection, whereas this thesis is focused on the detection of arguments in short chunks of text.

Inspired by the growing importance of social media's role in different aspects of everyday life, including social and political debates, this thesis uses Twitter as the main data source. Different statements have been collected and annotated on the geopolitical debate on Twitter about the expansion of the "Nord Stream" gas pipeline in northeast Europe, under the name "Nord Stream 2". The construction of the pipeline has primarily financial incentives, but it also raises concerns on political, environmental, and ethical issues, as there is a conflict of interest between parties and bodies that act for both national and European interests [42]. Concepts such as energy union and energy diversification are manipulative depending on the goal of each party and create strong arguments for or against the construction of the pipeline.

The lack of clear existing annotation guidelines creates a series of challenges in the design of the annotation process. First, the definitions for the different entities of argumentation provided in Chapter 3 offer a useful point of reference that guides the annotators, without restricting them to rigid norms. Then, the agreement level between the annotators had to be studied and improved until reaching an acceptable level. Multiple iterations took place between the annotators updating the annotation rules following an agile paradigm rather than a traditional top-down approach with rigid annotation rules. The iterations finalised the annotation guidelines, and they also contributed to the formation of the theoretical definitions presented in the previous chapter. This bidirectional approach between the definition of the different terms and the annotation process could be adopted in different unexplored NLP areas, providing a more robust theoretical foundation.

The majority of the existing research on argumentation detection does not prioritise the annotation strategy, but rather the modelling process and configuration of the ML algorithms that are followed for the classification task. This thesis is one of the few research items that provide details on the annotation process. The annotation methodology is in-line with the previous research approaches, establishing rules between the annotators and evaluating their agreement following the standardised processes. The dataset that has been created, and offered freely upon request, covers a previously unexplored topic that has gained great importance over the last year due to the increase in energy prices. The dataset has two versions, the first one includes the text body of the tweet and the final annotation verdict (argumentative, non-argumentative), the second version includes some additional information for the annotation such as the category the tweet falls into, and how the annotators labelled

each tweet.

In the first attempt of annotation, the annotators, the author alongside a second trained annotator, are asked to define a tweet as argumentative or not. The percentage agreement calculation was 0.2 and the Cohen's kappa equalled 0.0066. The results were disappointing, thus new rules were defined from the beginning after a conversation between the two annotators. A bottom-up approach was followed because it was easier to define specific attributes that are characterized as argumentative, even though ambiguous cases still exist.

In the bottom-up approach, it is asked from the annotators apart from characterizing a tweet as argumentative or not, to decide if the stance that exists in the tweet is explicit and in which subcategory the tweet belongs in. Five different sub-categories were defined: news title, reflection of feelings, publication of a report, expression of an open question, and, ironic statement. The results were significantly better in the second attempt, as the IAA were on 0.73 and the Cohen's kappa reached 0.45 that is considered moderate agreement [85].

A second discussion was held to improve the IAA even more. It was decided to concatenate the categories of report and title, and add the category of the cited opinion. Additionally, a list of words that indicate stance was created, it was decided that if the annotators are unsure about a tweet and it was difficult to understand, then it was classified as non-argumentative, the use of irony tends to indicate an implicit argument, and the appeal to an authority tends to indicate argumentative statement. The third annotation effort improved Cohen's kappa from 0.45 to 0.52, whereas Cohen's kappa for implicit/explicit agreement did not change significantly as from 0.33 it dropped to 0.32. Concerning the raw agreement to category agreement it reached 0.58 from the previous 0.36.

In the fourth annotation attempt, specific problems were founded and prioritized to increase the IAA in the argumentative / non-argumentative category. For that reason, more explicit rules were defined in cases where two or more categories are mixed. It was decided that the list of intense words would be categorized more descriptively, thus when they coexist with cited opinion/title more clear decisions can be made. Three different lists of words were created to aid the task of annotation. The first category lists words that are strong enough to express an argument without the need for supplementary information (e.g. trust, security, political), the second category lists words that need to be followed with some kind of backing a specific opinion (strategic, legislation, criticism), and the third category lists mostly transition words that could indicate argument, but they require strong backing (because, hence, thus). An example of an argumentative statement is a cited opinion combined with words from the second category. Concerning the implicit/explicit annotation, the words from the first category indicate explicit argument, whereas the words from the second and the third category sometimes indicate explicit and some other times implicit argument. No strict rules were defined concerning the number of words in the third category that might indicate explicit argument. As a general guideline, a cited opinion backed up by a word in the second category is characterized as explicit, whereas a cited opinion backed up by a word in the third category is characterized as implicit.

It has to be stated that even though the rules are defined in detail, it was expected from the annotators to face instances that do not belong in any of the aforementioned categories because statements in Twitter present great heterogeneity. Eventually, the dataset consists of 590 annotated statements and it has 452 statements annotated as non-argumentative and only 138 of them as argumentative. It is a highly skewed dataset that has opposite skewness

compared to the one observed in Dusmanu et al. [36], where the 77.3% of the collected tweets express an argumentative stance. The inter-annotator observed agreement reached 87.2%, whereas the unweighted Cohen's Kappa score is 0.64.

Example	**Annotation**
Read our in-depth weekly overview on European natural gas matters http:// buff.ly/1N5bBr0 #StateEnUn15 #Groningen #NordStream2	Non argumentative
Report: German Finance Minister teams up with Putin on #NordStream2 http:// buff.ly/1MQzItm	Non-argumentative
Deeply concerned about Germany's lack of concern with EU law for #NordStream2. RT if you are deeply concerned, too. @AuswaertigesAmt	Argumentative
What would be your question for high German official about #NordStream2, #energy policy #energyunion ? Thanks Twitter. :-)	Non-argumentative
U.S. LNG exports to EU, a powerful potential alternative to Russian-sourced energy. But is EU married to Gazprom? #GIPL #NordStream2	Argumentative

Table 4.3: Examples of Argumentative or Non-Argumentative

Providing some qualitative aspects of the dataset, the statements are not homogeneous but they express different perspectives on the discussion. Table 4.3 presents five examples of the collected tweets alongside their annotations offering an insight into the available information. The examples fall into different categories (news title, ironic tweet, open question) presenting examples from different perspectives. The number of tweets (∼45000) that have been written and the number of user accounts (∼7000) that have mentioned the construction of the Nord Stream2 pipeline reveal the great interest of the audience for this transnational debate.

The iterations of the annotation process had continuous interaction with the development of the theoretical definitions as expressed in the section *3.4 Argumentation modelling*, revealing a bidirectional relation. First, an outline with the definitions for the terms that are used was provided, which was used as a point of reference for the first version of the annotation process. Then, in every annotation cycle, the definitions were re-visited, adjusted, and improved by removing any ambiguity, finally guiding the annotators to better classification decisions for the different textual segments.

More specifically, *Definition 1* was provided in the beginning of the annotation process, and it was used to present the consensus to the problem. During the annotation process, *Definition 2, 3,* and *4*, for topic, claim, and reason, respectively, were used as guidelines helping the annotators throughout the process. The definitions were formed before the beginning of the annotation process and they had a significant role in it, while they were improved based on the feedback that was received. Similarly, *Definition 5* and *6* were used to assist the annotators to lead in the final decision to ensure they follow the user's train of thought and classify the statement as an argument or not. The decision was confirmed by following the insight provided in *Lemma 1.* Finally, *Definition 7* provides a synopsis for the main terms that are used in argumentation through a framework that can be applied on both manual and computer-aided annotation.

An example of the annotation process' flow for the first argumentative statement given in Table 4.3 *"Deeply concerned about Germany's lack of concern with EU law for #NordStream2. RT if you are deeply concerned, too. @AuswaertigesAmt"* follows. First, the topic of the statement is defined by identifying the hashtag *#NordStream2*, the usual navigation method

over Twitter. The claim is identified via the use of verb *concern* while the reason is *Germany's lack of concern with EU law*. The tweet also initiates a discussion with other users, but it does not affect the annotation process. The reason can easily map with the claim, since the possible bypass of EU law from a country-member could create political turmoil in the entire area. Consequently, the clear expression of the claim is supported by a valid reason and leads to the creation of a valid reason according to *Lemma 1*.

4.3 Discussion

Modern AM systems usually need to address many issues including argument detection, reasoning evaluation and source identification. In addition, social media have transformed the means of communication and information exchange, promoting shorter bursts of text and fluid argumentation structures having problems controlling the spread of rumours and fake news. In this noisy environment, it is important to develop mechanisms that are capable of identifying argumentation in short text revealing previously unexplored capabilities in the wider spectrum of the NLP domain.

The emerge of Web 2.0 has provided tremendous potential in the NLP research community as it offers an endless and free source of data covering a wide spectrum of topics. When researchers choose social media as their data source, they tend to prefer political and social debates such as Brexit, national elections, constitution changes, etc. In this work, a dataset on the debate for the construction of the natural gas pipeline Nord Stream 2 is created. A topic that apart from its economic impact, has also emerged as political debate in the European Union (EU).

The annotation process is an important part of every research on the NLP domain. More specifically, when the research takes place in real-life settings the challenges that are found should be stressed and addressed in order for best practices to be extracted. The difficulties that occur during the annotation process on a dataset that represents real-life settings indicate the challenges that will be faced later on in the application of theoretical frameworks for any NLP task. Therefore, it is important for every theoretical framework to be applied to more than one dataset, preferably from different sources. In this thesis, the proposed framework is tested on two different datasets, the first one comes from a debate on Twitter and the second one are transcripts from a political debate for presidential elections.

The annotation process for the task of argumentation detection followed the theoretical modelling and the AFAD as described in chapter 3, assuring its scientific integrity. The lack of established theoretical foundations created a bidirectional relation between the annotation process and the finalization of the definitions. The annotation method can be applied to argumentation detection tasks in different use cases, while it is well enough documented to be applied on different NLP tasks as long as the necessary adjustments take place. The same approach, with minor alterations, can be applied to different AM-related tasks following an iterative process between the definition of terms and the annotation of the dataset. The iterations would be really useful in the recently emerged NLP tasks whose terms are not yet finalised, thus creating theoretical definitions and terms that are tested in real-life settings.

Table 4.4 presents the characteristics of the datasets that will be used later in the thesis. The first column has the name of the datasets, the second one the percentage of the argumentative statements, the third and fourth columns the size and the source of the

dataset, respectively, and the last column the IAA calculated with Cohen's kappa. Nord Stream 2 and Miliband datasets are heavily skewed towards non-argumentative statements, whereas Cameron and Clegg are more balanced between the two classes. However, the non-argumentative class still outnumbers the argumentative one illustrating that in real-life settings argumentative clauses are not the dominant form of expression.

Dataset	**Arg. pct.**	**Size**	**Source**	**IAA**
Nord Stream 2	23.4 %	590	Twitter	0.64
Cameron	46.7 %	122	Speech	0.52
Clegg	44.2 %	144	Speech	0.57
Miliband	30.8 %	160	Speech	0.47

Table 4.4: Details on the different datasets that have been used in this thesis.

Recalling the initial research questions that have been expressed at the beginning of the thesis, this chapter reveals the connection between argumentation detection and modern NLP applications showing how related studies connect different NLP tasks (research question 1). Additionally, it covers the second contribution as stated in the introduction, illustrating the heterogeneity in argumentation techniques in Twitter and presenting the annotation process for a previously unexplored topic, at least to my knowledge. The constructed dataset will be available upon request for the following 12 months, and then it will be published on an open repository. In the next chapter, the first case study is presented, deploying different ML algorithms, and testing the proposed argumentation model using the Nord Stream 2 dataset.

Chapter 5

Argumentation detection in social media

The growth of social media in the last decade has created new means and codes of communication disrupting the traditional ones such as newspapers, journals, and books. Even though multimedia channels, like vlogs and podcasts, have started to emerge, they have not yet overrun textual-based social media like Twitter and Reddit which are often used by researchers as a textual data source. The surge in Web-generated data has maximized the potential of the NLP research community since it offers an endless free source of data covering a wide spectrum of topics. When researchers choose social media as their data source, they tend to prefer political and social debates such as Brexit, national elections, and gay marriage to include aspects of social value [114, 65]. However, even carefully selected online resources include noise and do not follow traditional argumentation techniques. The challenges that have appeared require both theoretical modelling schemes and the application of novel methods. Modern research problems, such as intelligent information retrieval systems, must be capable of understanding not only what people think, but also why people take a stance on a given issue. The third chapter of the thesis provided the theoretical foundations of argumentation modelling in short text, while this chapter presents different implementations and provides a more practical perspective.

The rest of this chapter is organised as follows. In section 5.1 a short introduction takes place illustrating the challenges of the task of argumentation detection in social media. Then, section 5.2 presents the three different approaches that have been followed to answer the task at hand, rule-based implementation based on the AFAD, machine learning algorithms, and a hybrid solution. The results are presented in section 5.3, and, finally, section 5.4 initiates the discussion on the impact and the potential of argumentation detection in social media.

5.1 Introduction

Argumentation detection is often neglected in numerous AM pipelines when they are applied in settings where the existence of argumentation is given, such as persuasive essays or rhetorics competitions. In the last decade, the means for communicating and expressing arguments have changed. The extensive use of the Web has provided the capability to express arguments and experiences on different online platforms. In the mid-2000's user-generated content has

started to gain a significant role in the design and implementation of websites, realizing the concept of Web 2.0. Looking at 2020 and beyond, social media have, and will continue to have, a dominant role in every social debate.

Social media can be characterised as an extension of Web 2.0 allowing not only the communication between the Web and the user but also facilitating the communication between the users. The rise and extensive use of social media in the last decade have established them as both news and communication platform. The potential of the users' influence in the diffusion of news and shaping the public opinion has gained the interest of the research community from different fields, studying both social and technical aspects of this new means of communication. The speed of exchanging information and the comfort in their use has enabled social media such as Facebook, Twitter, and Reddit to replace, to some extent, traditional means such as newspapers and television.

Twitter has gained the greatest research interest from the NLP community among the existing social media platforms, mostly due to its format that allows users to express their opinions and personal experience on a variety of topics from political debates to fashion trends. Twitter is a social networking and micro-blogging service, enabling registered users to read and post short messages, so-called tweets [61, 169]. The inherited public nature of tweets allows users to interact with multiple topics and users in real-time without any significant constraints, providing a previously unexplored field of application for the NLP community. The upcoming challenges require a need to develop innovative research methods and mechanisms that enhance the trust and the explainability of the AI services and enable the transferability of knowledge between tasks and domains. In the meantime, the research questions that have emerged are now more complex and require high levels of intelligence and an in-depth understanding of the human language.

Other popular social media platforms such as Facebook and Instagram, even though they offer capabilities for textual interaction between users, visual representations play a crucial role and they should be included in the analysis, hence they are not extensively used in the NLP domain. Reddit is an alternative social media platform that could be used as a primary data source [162], but its inherited anonymity removes credibility from its users [19].

In contrast to Reddit's community, political leaders and activists around the world maintain Twitter accounts with active presence [147]. They update and interact with their followers providing ample opportunities for research in the field since political and social debates attract the interest of the audience. Therefore, when researchers choose social media as their data source, they tend to prefer political and social debates. The first case study of the thesis utilises online data from Twitter regarding the construction of the natural gas pipeline Nord Stream 2, a public debate with great interest among members of the EU.

This chapter covers the emerging area of argumentation in Twitter, in an attempt to quantify previously unexplored qualitative aspects of the language. In summary, the main contributions of this chapter are:

- Demonstrates different proof-of-concept implementations for the AFAD.
- Confirms the added value of integrating rule-based mechanisms into ML algorithms on highly imbalanced datasets.
- Presents the performance of four different ML algorithms and the effect of applying additional features to them.

The rest of the chapter is organized as follows: Section 2 presents the experiments that took place including the implementation of the rule-based approach, the deployment of the ML algorithms, and the proposed hybrid solution. Section 3 presents the results of the experiments. Section 4 initiates a discussion and presents the challenges.

5.2 Methodology

Related work on AM in the short text (the predominant source is Twitter) usually includes the task of argumentation detection as a necessary preliminary step in their designed pipelines, even though they have different objectives. Argumentation detection is used as a means of achieving goals such as evidence recognition [1], source identification [36], and construction of argumentation graphs [17]. The aforementioned studies follow a data-driven approach, which, although successful and easy to use, does not promote trust and explainability in the AI methods. An alternative is the adoption of a rule-based approach using semantic proximity between known claims and previously unknown tweets [187].

In this section, three different approaches are implemented, presented, and compared. The rule-based mechanism, subsection 5.2.1, offers an implementation of the AFAD, subsection 5.2.2 presents the implementation of different ML algorithms, and, finally, subsection 5.2.3 proposes a hybrid solution for argumentation detection enhancing the trust in the outcome of the ML algorithms.

5.2.1 Rule-based mechanism

The development of argument base (AB) is based on the idea of collecting arguments and relations, where each argument formalises knowledge conducive to solving the problem in question. This knowledge is used to formalise domain knowledge and develop symbolic AI algorithms that are capable of identifying argumentation in a short statement. Combining this approach with the provided definition of argument in short text, the AB is separated into two entities, claim database (claim DB) and reason database (reason DB). In the meantime, any potential relations are disregarded, while focusing on the detection of argumentative text. The claim DB includes common claims that are expressed in the debate and the reason DB contains the reasoning that supports the claims.

The rule-based approach introduces a method capable of identifying the conceptual proximity between previously unseen chunks of text and the collection of claims/reasons. For example, the tweet *"EU should block #NordStream2 on climate grounds"* and the claim *"the gas pollution is the main threat for the biodiversity of the Baltic sea"* are correlated, even though they do not share any common words. For the construction of the claim DB and the reason DB, claims and reasons are manually extracted during the annotation task and they express the main objectives of the public debate. Consequently, the knowledge stored in the AB is domain-dependent and cannot easily be transferred to new domains.

More specifically, the AB for the #NordStream2 debate was created from the two annotators during the iterative process of the annotation. In every iteration, the annotators tracked down the claims and the reasons for the argumentative tweets and at the end of every iteration, there was a discussion on the validity of each pair of claim/reason. Due to the nature of the topic, a technical project with political and environmental implications that

attracts the interest of a wider audience, it was challenging to group the arguments under specific labels.

Table 5.1 provides six examples of arguments split into claims and reasons providing some examples of the nature of the arguments that are used. The table presents the arguments as expressed in Twitter in a sentence format, whereas in the algorithm only uni-grams and bi-grams were used. For example, the hashtag #NordStream2 was rejected because it is used to define the topic and it cannot be used as a claim. Additionally, due to the need to express claims and reasons as uni-grams or bi-grams, in the first example of the table, the reason is split into three autonomous reasons expressed from the different hashtags: #climate, #EnergyUnion, #ParisAgreement.

Claim	**Reason**
block #NordStream2	#climate grounds #EnergyUnion #ParisAgreement
block #NordStream2	Security grounds
#NordStream2 can wait	undermine EU policy goals
support #NordStream2	contributes to all 3 key objectives
support #NordStream2	promotes further integration of energy market
support #NordStream2	clear economic benefits

Table 5.1: Examples of claims/reasons pairs that are used to construct the AB.

The correlation between a tweet and the collected claims/reasons (mapping functions) can be estimated in multiple ways. However, a sequence of symbols has limited usefulness as they cannot be fed directly to many algorithms because most of them expect numerical feature vectors with a fixed size and not a series of characters with variable length. For this reason, a processing pipeline has been designed. The very first step of the pipeline is a pre-process function that eliminates the evident noise such as stopwords, URLs, leading and trailing whitespaces. More specifically, the value of URLs is limited since they re-direct to external resources, typos such as whitespaces are not often in the context of a political debate and they do not offer any semantic value. Finally, the stopwords could have semantic value in a specific context (e.g. "we should not do this"), but after a preliminary analysis, they have been disregarded because it was too complex to express valid argument only with the use of stopwords. On the other hand, hashtags and mentions are striped (e.g. #NordStream became NordStream), because hashtags can contain claims or reasons (e.g. #climatechange), while mentions can be used as citations to authority. The tokens that have been collected construct the dictionary which is expanded whenever a new entry is introduced in the dataset. Eventually, a corpus of phrases/sentences is represented by a matrix having the tweets of the dataset as rows and the tokens of the dictionary as columns.

The process of turning a collection of phrases/sentences into numerical feature vectors is called vectorization, and there are different ways of expressing the relation of a phrase/sentence with a concept. The most straightforward solution, term frequency (TF), counts the recurrences of each token while completely ignoring the relative position information of the words in the corpus. This method could lead to bias errors due to the repeatability of specific terms in a specific domain without any significance. A technique for weighting different terms in the vectorization process is the term frequency-inverse document frequency

(TF-IDF). The first part is the number of occurrences of the term in the statement, and the second part is the inverse of the number of statements that contain the specific term. Therefore, the specificity of a term can be quantified as an inverse function of the number of statements in which it occurs. For this research study, both approaches have been followed and a series of comparisons are presented.

However, the challenge lies not only in the transformation of the text into a binary format that a machine can understand but also in measuring the semantic similarity of the constructed vectors. A way of estimating the conceptual proximity between a tweet in the dataset and an argument in the AB is the calculation of the cosine similarity between two vectors. Let t and a be two vectors of the same size representing a tweet in the dataset and an argument in the AB, respectively. Using the cosine measure as a similarity function, we have:

$$sim(\mathbf{t}, \mathbf{a}) = \frac{\mathbf{t} \cdot \mathbf{a}}{\|\mathbf{t}\|\|\mathbf{a}\|} \tag{5.1}$$

where $\cdot$ is the dot product of the two vectors, and $||\mathbf{t}||$ is the Euclidean norm of vector $\mathbf{t} = (t_1, t_2, ..., t_n)$, defined as $\sqrt{(t_1^2 + t_2^2 + \ldots + t_n^2)}$. Similarly, $||\mathbf{a}||$ is the Euclidean norm of vector $\mathbf{a}$. The multiplication of the two vectors is achieved using the transpose operator allowing us to multiply the components of the two vectors by flipping the transposed vector:

$$\mathbf{ta^T} = \begin{bmatrix} t_1, t_2, ..., t_4 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_4 \end{bmatrix} = t_1a_1 + t_2a_2 + ... + t_na_n \tag{5.2}$$

The measure computes the cosine of the angle between vectors t and a, returning values between [0, 1]. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal), thus they do not have any semantic correlation. On the other hand, the closer the cosine value is to 1, the smaller the angle and the greater the correlation between the two vectors. Negative values do not exist because the encoding methods we have selected do not assign negative values to the text, however, a different encoding method could produce negative values.

5.2.2 Machine learning algorithms

Machine learning classification algorithms automatically classify a set of previously "unseen" text segments to a set of predefined class labels based on previously labelled data on which the algorithms have been trained. This process requires resources that are related to the given task as well as useful features that can be extracted from either plain text or metadata. Social media are typical resources that can provide a large amount of user-generated data with semi-structured or structured metadata, such as geo-tagging, date of creation, user ID, etc. The same pre-process function that has been deployed in the rule-based component has also been applied here, removing stopwords, URLs, leading and trailing whitespaces, and striping hashtags and mentions. A wide range of statistical and linguistic features have been suggested for argumentation detection and other NLP tasks, such as sentiment analysis and source identification. In this research, a wide set of features has been chosen covering the following categories:

Lexical features refer to the n-gram encoding technique (uni-gram, bi-gram, etc.). Uni-gram and bi-gram encoding techniques have been deployed with both TF and TF-IDF techniques, creating four comparable clusters. This variety in encoding techniques provides an extensive overview of their applicability for the task of argumentation detection.

Semantic features define the characteristics of the language that can provide a deeper insight into the data such as Part-of-Speech (PoS), dependency relations syntactic and parse trees. The NLTK PoS tagger [12] has been used which can identify and group words into different categories that display similar syntactic behaviour.

Sentiment features reveal emotions and they are usually detected with the use of external lexicons. The textblob software [165] has been used for the extraction of two features: polarity and subjectivity. The former returns a float number within the range [-1.0, 1.0], and the latter within the range [0.0, 1.0]. The degree of polarity can be interpreted as a negative/positive stance towards a specific topic, whereas a high subjectivity score correlates with opinionated claims.

Twitter-specific features are offered as metadata through the Twitter API and concern the specific characteristics a tweet has, such as the length of a message, the presence or not of URLs, mentions of other users, hashtags, and official account verification. Based on these characteristics, binary variables have been used indicating the existence of mentions and hashtags, as well as a counter variable for the characteristics.

Algorithm 1: Execution of ML Algorithms

```
Result: A list of tuples;
Define n for text n-gram encoding;
Define list of external features;
Define list of algorithms;
Pre-process the dataset;
while i ≤ n do
    Encode dataset with n-gram;
    foreach algorithm in algorithm list do
        foreach combination in feature list do
            Execute algorithm with combination in dataset;
            Cross-validate class prediction;
            Store the results;
        end
    end
    Increase the n;
end
```

Four ML classifiers have been trained for the task of argumentation detection with and without the use of the aforementioned features. Algorithm 1 illustrates the execution of the designed pipeline. For the execution of the ML algorithms, the encoding mechanism (uni-gram, bi-gram, etc.) has been provided as an input parameter (defined by the variable n), and the option for the use of external features has also been given. After a preliminary analysis, the list of features has been limited to the use or not of all the external features. Eventually, the iterative execution of the ML algorithms returns a list of results based on

the different combinations that have been performed. Due to the limited size of the dataset, cross-validation has been used to prevent over-fitting errors and provide higher reliability in the results.

The algorithms which have been selected to be used in this research represent different algorithmic approaches. In total, four different ML algorithms were deployed and executed providing a wide test area for their performance in the specific task. The chosen algorithms with a short description follow:

Parameter	**Value**	**Short description**
hidden layer sizes	(100,1)	Number of hidden layers(1) and units(100)
activation function	ReLU	Piecewise linear function defining the output of each unit
output function	Logistic	Logistic sigmoid function defining the output of the last layer
solver	Adam	Stochastic gradient optimization algorithm
L2 penalty	0.0001	Weight penalty set to Adam solver to avoid overfitting
batch size	200	Size of batches used in Adam solver
learning rate	constant	Fixed learning rate
learning rate init	0.001	The step-size for weights update
max_iter	200	Number of epochs for Adam solver
shuffle	True	Shuffle samples in each iteration
random state	0	Define random number generator for reproducible results
tolerance	1e-4	Minimum acceptable change in loss or score
n_iter_n_change	10	Maximum number of epochs to not meet tolerance improvement
beta_1	0.9	Exponential decay rate for the 1st moment
beta_2	0.999	Exponential decay rate for the 2nd moment
epsilon	1e-8	Stabilizer value to prevent any division by zero

Table 5.2: The parameters that have been used for the deployment of the MLP.

Multi-layer Perceptron (MLP) belongs to the family of the artificial neural networks (ANN) and it is based on a function $f(\cdot) : R^m \rightarrow R^o$ that is trained on a given dataset, where m is the number of dimensions (features) for input and o is the number of dimensions for output (two alternatives in the argumentation detection task). The transformation of the input to the desired outcome is realised through an activation function. In this research work, the Rectified Linear Unit (ReLU) has been deployed in the hidden layers and the logistic sigmoid function in the output layer. The ReLu has been selected due to its non-saturation capabilities and its low computational requirements while the sigmoid function is due to the binary nature of the problem. The two main disadvantages of the ANN algorithms are the difficulty of the fine-tunning process and the often unexplained behaviour of the network due to its complexity, which is calculated as $O(n * m * h^k * o * i)$ where n is the training samples, m the features, k the hidden layers, h the neurons, and o the output neurons. In the designed architecture for the Nord Stream 2 dataset, there are 531 training samples, 8 features from the Twitter metadata, 36 features from the POS tagging, plus the number of the n-grams that are formatted for every sample; there is one hidden layer with 100 neurons, and two potential outcomes. Additionally, the iterations due to the k-fold cross validation increase the calculation complexity because the training set is split into 10 sub-totals. The total number of connections could create some obstacles to scaling the problem due to its processing demand unless some parallelization techniques apply. Table 5.2 provides the value

of the MLP parameters that are used in the Nord Stream 2 use case.

Decision Tree (DT) is a non-parametric supervised learning method that predicts the values of the required feature through a series of decision rules inferred from the dataset's features. Due to the design of the decision rules, it is capable **of handling** features that indicate categorical data such as the existence or not of specific hashtags or mentions. Another advantage of the DT solutions is the easy interpretation of the algorithm's flow. The DT presents complexity of O(n*d*log(n)) where n is the number of samples, d is the number of features, and *log(n)* is the depth of the tree. On the negative side, DT tends to present a skewness towards the majority class, stuck in local optima and overfit on a dataset with many features. Therefore, its performance on imbalanced data for complicated problems could be questionable while it would be challenging to apply it on larger datasets. Table 5.3 presents the values that are given for the implementation of the algorithms with a short description of their operation. The Gini criterion has been used to evaluate the quality of each split following the best strategy without constraints on the maximum depth of the tree. The number of features that are used depends on the number of unigrams of each statement, hence it is estimated to be around 45 features.

Parameter	**Value**	**Short description**
criterion	gini	Function evaluates the quality of the split
splitter	best	Strategy to split at each node
max_depth	None	Maximum depth of the tree
min_samples_split	2	Minimum samples to split an internal node
min_samples_leaf	1	Minimum samples to create a leaf node
min_weight_fraction_leaf	0.0	Minimum weight fraction of samples to create a leaf node
max_features	45	Number of features to decide the best split
random_state	0	Randomness of the estimator
max_leaf_nodes	None	Number of leaf nodes
min_impurity_decrease	0	Minimum number of impurity to split the node
class_weight	None	Weights associated with classes
ccp_alphanon	0	Complexity parameter used for Minimal Cost-Complexity Pruning

Table 5.3: The parameters that have been used for the deployment of the DT.

Logistic Regression (LR) is a statistical classification algorithm that is used to model the class probability of the required feature and then convert the probability to log-odds through the logistic function. The logistic function that was used for this research is the Limited-memory BFGS (L-BFGS) from the family of quasi-Newton methods [20]. The *L2* penalty for the regularization has been selected, the inverse of the regularization strength is set to 1.0, the primal formulation has been selected because the number of features is not greater than the number of samples, the tolerance for stopping criteria is set to 1e-4, the constant values are included in the decision of the algorithm, and the two classes have been assigned with the same weight. Overall, LR is considered an interpretable and computationally inexpensive presenting *O(nd)* where n is the number of samples and d the features that are used for this use case. On the negative side, the LR struggles to solve non-linear problems. Therefore, its application in a noisy environment could be questionable and different solvers with different parameters should be tested in different use cases.

Support Vector Machines (SVMs) are a set of supervised learning methods that rep-

resent the samples as points in space and through a mapping process, the points are assigned into classes, producing a non-probabilistic binary linear classifier. Moreover, SVMs can also perform non-linear classification using different kernel functions which map the input into high-dimensional feature spaces. However, the number of features that have been extracted for this research work is large, hence there is no need to map data to a higher dimensional space and the linear kernel has been used. Providing more details on the implementation of the SVM, the regularization parameter of the squared L2 penalty has been set to 1, the shrinking heuristic has been enabled for shorter training time, the tolerance for stopping criterion is set to 1e-3, the cache size has been set to 200 MB, the two classes have been assigned to have the same weight, and finally, no limit has been set on iterations within the solver. Another characteristic of the SVMs algorithms is that they need a clear margin of separation between classes to outperform other solutions; a prerequisite that is tough to be met on complex or vague classification tasks.

5.2.3 Hybrid approach for argument detection

In symbolic AI a set of rules are applied to transform input into knowledge presenting satisfactory results for problems with clear cut settings. The explainability in their performance and the accountability in their design process, set symbolic AI as a valuable alternative in different use cases that require special data handling, such as medical diagnosis [5, 31]. However, symbolic AI models use hardcoded knowledge and rules to tackle specific problems, hence they only work in very narrow use cases. Whenever the problem is generalised there is the need for adding new rules to cover previously unknown conditions, thus their use in social media text with constantly changing parameters is questionable.

The problem of scalability is encountered in data-driven solutions by employing ML algorithms. The ML algorithm can more easily deal with messy and unstructured data without requiring the creation of specific rules from domain experts. In the last decade, ML algorithms have dominated the field of NLP due to the large availability of data and the easy accessibility to processing power presenting impressive results in different tasks. However, recent objections and criticism on the ethical use of data, biases, and the overall obscure decision process of these algorithms, have raised some concerns on aspects such as trust, accountability, and explainability [172, 11].

The objective of the proposed hybrid methodology is to enhance the weak aspects of ML through the modification of a supervised ML pipeline capable of fixing possible bias of the algorithms towards the majority class. The combination of the rule and ML-based approaches has the potential to create a hybrid solution capable of integrating the positive aspects of each one. Under this rationale, the designed pipeline forwards the predictions of the ML algorithms to the rule-based component which is responsible of fixing any anomalies that are identified.

The proposed architecture is depicted in Figure 5.1 illustrating the combination of the ML and rule-based approaches. The ML component receives as input the extracted features (lexical, semantic, sentiment, Twitter-specific) from Twitter, which are used as the main data source. Afterwards, the ML algorithms process the collected data trying to find patterns and eventually determine the class (argumentative / non-argumentative) for each sample. For the rule-based approach, the claim DB and the reason DB are normalized into a common format, and alongside Twitter data are fed into the vectorization algorithm. The predictions of the

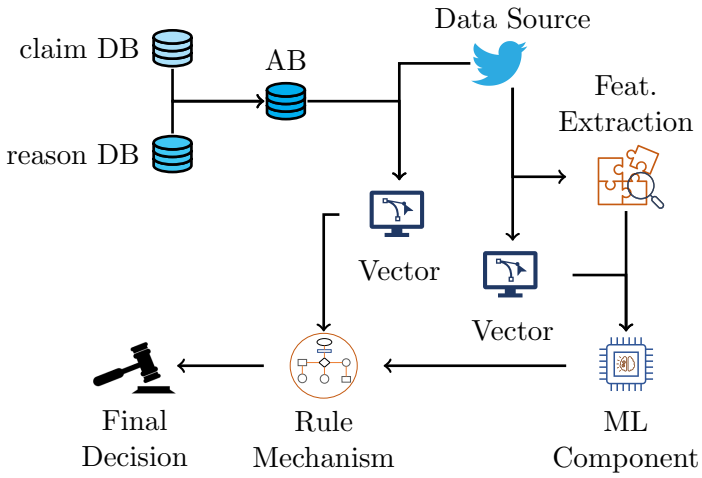

Figure 5.1: A graphical illustration of the hybrid approach that has been followed.

ML components are compared to solutions suggested by the rule-based mechanism and when there is a controversy between the two predictions, the decision mechanism is enabled. The decision mechanism is based on the idea that the ML algorithms tend to underestimate the minority class while the rule-based mechanism is built to identify even implicit arguments. In terms of performance, it translates into high precision for ML algorithms and high recall for rule-based algorithms. For highly negatively imbalanced datasets, such as the Nord Stream 2 dataset, it is important to increase the correct prediction of the positive instance, even if it means decreased performance in the majority class. Therefore, the hybrid decision mechanism keeps the positive ML predictions while in other cases, the rule-based suggestions are preferred due to the domain knowledge that carries.

Algorithm 2 illustrates the hybrid solution utilizing the concept for both semantic similarity and ML. The algorithm receives as input: 1) the tuple of text/prediction from Algorithm 1, and 2) the constructed AB that contains domain knowledge. Then, through three consecutive loops, the semantic similarity between text/claims, and text/reasons are estimated. The two similarity scores are added up producing the total semantic similarity score between the tweets and the AB. Finally, the median score is calculated, and the statements that are above this threshold are classified as argumentative, and those that are below are classified as non-argumentative.

An alternative hybrid implementation would integrate the similarity score of the rule-based methodology as additional features, but a preliminary analysis presented poor results, hence it is not been preferred. Finally, we should consider that in the implemented method the sensitivity of the rule-based mechanism can be easily tuned, thus any potential adjustments can remain under control. For this research study, the rules have been designed to identify the instances in the minority class and therefore increased recall is expected.

5.3 Results

The previous chapter presented the different approaches that have been implemented and this one presents their results. The performance of the different solutions is reported using

Algorithm 2: Execution of hybrid solution

```
Result: A list of tuples
Receive tuple texts/predictions from Algorithm 1;
Receive AB;
Vectorize text and AB;
foreach text-prediction in tuple texts/predictions do
    foreach claim in claim DB do
        foreach reason in reason DB do
            Find semantic similarity between text and claim;
            Find semantic similarity between text and reason;
            Estimate total semantic similarity;
            Correlate total semantic similarity with prediction;
            Draw final decision;
        end
    end
end
```

different metrics to provide a good overview of the algorithm's performance. Subsection 5.3.1 presents the performance of the rule-based approach, subsection 5.3.2 reports the results from the ML algorithms, and subsection 5.3.3 presents the results of the hybrid solution.

5.3.1 Rule-based approach

For the first set of experiments, the rule-based approach as described in subsection 5.2.1 was implemented. They present different implementations for the functions $g(r_i, t)$ and $h(c_j, t)$ representing the mapping from reasons to claims and from claims to arguments, respectively. The first scoring function, $sem\_AB(x)$ function, consists of two distinct tasks; the first computes the semantic proximity between the text and the set of reasons and the second one computes the semantic proximity between the text and a set of claims. Both sets have been created manually by the author and express -to some extent- the spirit of the debate. The $sem\_AB(x)\_idf$ function follows the same principles but the encoding of the text takes place using the TF-IDF method. The third alternative uses an external general-purpose dictionary [165] that assigns values within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. After the annotation process, it has been observed that tweets without any subjectivity aspects are usually news titles that redirect to external resources, hence higher subjectivity is correlated with a higher probability of expressing arguments. Finally, for the last approach, a combination of the two methods was implemented ($comb(x)$ function), where a chunk of text has to display both semantic similarity with a known argument and reach a relatively high subjectivity score (aka above the median). Apart from the different functions, two more methods were also used as benchmarks. The first benchmark is a random function, and the second one is the Jaccard similarity coefficient score (or Jaccard index). The Jaccard index is used for gauging the similarity and diversity of sample sets, and it is defined as the intersection between the arguments stored in the AB and the collected tweets divided by the size of the union of the sample sets: $J(AB_i, T_j) = \frac{|AB_i \cap T_j|}{|AB_i \cup T_j|}$.

The use of baselines and, in general, methods that do not project to surpass more complicated approaches is a good point of reference for comparisons and it is used extensively in different tasks and environments. In the work of Sendi and Latiri [154] an evaluation formula for the task information retrieval using the count and number of words in both document and the collection is proposed. The formula follows:

$$c(w; D) = count\ of\ word\ in\ the\ document$$
$$c(w; C) = count\ of\ word\ in\ the\ collection$$
$$|D| = number\ of\ words\ in\ the\ document$$
$$|C| = number\ of\ words\ in\ the\ collection$$
$$numerator = c(w; D) + mu * c(w; C)/|C|$$
$$denominator = |D| + mu$$
$$score = log(numerator/denominator)$$

where the *mu* has been set to 2500, indicating that it cannot be used in small documents because the score differences would be very small. The formula is certainly a more advanced baseline compared to the random assignment which is implemented in the Nord Stream 2 use case while it slightly resembles the Jaccard similarity which is also presented to offer a more complete overview of the performance of the rule-based approach. Due to the dominance of the data-driven approaches, the definition of a rule-based method as the system's baseline is rare, instead, a ML algorithm with basic lexical features is usually considered the standard baseline. Some examples of baselines that are used for the task of argument detection are unigram encoding as input to three different ML algorithms (DT, SVM, Naive Bayes) [1], raw character counts and bi-grams in a logistic regression model [18], and, unigram, bigram, and WordNet verb synsets as inputs to DT and LR [36]. The following subsection, which presents the results of the ML algorithms, uses both unigram and bigram encoding without the inclusion of any external features offering the standard point of comparison for future research. The random baseline has been selected due to the heavily negatively imbalanced nature of the dataset and it presents the same F1-score with the execution of the DT with unigram input. However, in terms of F1-score with macro calculation, the random baseline performed poorly.

	Prec	Rec	F1 score	F1 (micro)	F1 (macro)
Baseline	0.25	0.56	0.35	0.53	0.49
Jaccard Sim.	0.28	0.64	0.39	0.56	0.52
sem_AB(x)	0.34	0.77	0.47	0.62	0.59
sem_AB(x)_idf	0.33	0.73	0.45	0.61	0.57
sub(x)	0.27	0.61	0.38	0.55	0.51
comb(x)	0.37	0.45	0.41	0.70	0.61

Table 5.4: Comparison of the different rule-based mechanisms that have been applied using three different techniques for estimating the F1-score.

Table 5.4 presents the results produced after the deployment of the six different rule-based techniques. The first column has the name of the deployed technique and the other five columns have precision, recall and F1-score calculated with three different methods. For the identification of the argumentative tweets, the $sem\_AB(x)$ function surpasses the rest of the methods with a 0.47 f1-score, followed by the $sem\_AB(x)\_idf$ with 0.45. The general-purpose lexicon comes up short on every calculation technique, but it is better than the random baseline, with 0.38 compared to 0.35 f1-score. The $comb(x)$ function fails to incorporate their benefits and presents the highest precision, but also a great drop in the recall. Finally, the Jaccard index comes up short with a 0.39 f1-score, suggesting that argumentation detection is a complicated task that requires more sophisticated approaches, however, it outperforms the random baseline indicating its suitability as an advanced baseline. All the algorithms underperform in terms of precision, with the $comb(x)$ function achieving 0.37 precision, which is the highest score. The $comb(x)$ function is expected to have the highest precision because it uses double filtering from two different methods, while it also receives the penalty in the recall having 0.45, even lower than the random baseline. On the other hand, in terms of recall, the rule-based algorithms present satisfactory results with the $sem\_AB(x)$ achieving the highest score with 0.77 recall. The rule-based algorithms are expected to underperform in terms of precision since they have been designed to identify even implicit arguments in a highly negatively imbalanced dataset, while the recall is expected to present satisfactory results. The trade-off between precision and recall can be adjusted by modifying the AB.

5.3.2 ML-based approach

The second set of experiments includes the execution of four different ML algorithms with different encoding mechanisms (TF and TF-IDF) on both uni-gram and bi-gram levels. Moreover, the algorithms are executed with and without additional features (semantic, sentiment, Twitter-specific) assessing the capability of the algorithms to exploit these extra features. The results provide a good overview of the performance of different ML algorithms for the task of argumentation detection.

Table 5.5 presents the F1-score calculated with binary, micro, and macro calculation for four different algorithms. The best performance is achieved when the MLP is deployed using TF uni-gram encoding with external features reaching 0.50 f1-binary and 0.69 f1-macro score. The second-best performance is observed when the SVM is executed with TF uni-gram encoding while using external features, presenting 0.47 f1-binary and 0.67 f1-macro score. In terms of micro calculation, the performance of the algorithms is more than satisfactory having the lowest score of 0.73 f1-micro, and the highest of 0.81 f1-micro. The impact of the encoding method seems to be the most important factor since every algorithm presents -almost- always better results when they receive text encoded with the TF technique. Additionally, the use of bi-gram does not seem to offer any additional value in the classification task. The limited length of the text and the protection of the special characteristics of Twitter (e.g. hashtags, mentions) sets the use of TF-IDF encoding ineffective. Regarding the use of additional features, even though the highest and the second-highest score is achieved using additional features, no clear conclusion can be drawn. It seems that the use of excessively complicated encoding does not provide the expected results for the task of argumentation detection.

			F1-binary	F1-micro	F1-macro	F1-binary	F1-micro	F1-macro
N-gram			Encoding			Encoding + Features		
Uni-gram	TF	MLP	0.45	0.81	0.67	0.50	0.80	0.69
		DT	0.35	0.76	0.60	0.28	0.75	0.56
		LR	0.38	0.80	0.63	0.40	0.81	0.64
		SVM	0.44	0.79	0.65	0.47	0.79	0.67
	TFIDF	MLP	0.46	0.80	0.67	0.46	0.79	0.67
		DT	0.31	0.75	0.58	0.27	0.73	0.55
		LR	0.00	0.77	0.44	0.20	0.79	0.54
		SVM	0.25	0.80	0.57	0.29	0.79	0.58
Bi-gram	TF	MLP	0.41	0.81	0.65	0.26	0.80	0.58
		DT	0.30	0.75	0.58	0.26	0.75	0.55
		LR	0.28	0.80	0.58	0.34	0.81	0.61
		SVM	0.43	0.82	0.66	0.42	0.81	0.65
	TFIDF	MLP	0.43	0.80	0.65	0.41	0.81	0.65
		DT	0.35	0.76	0.60	0.27	0.74	0.56
		LR	0.00	0.78	0.44	0.13	0.78	0.50
		SVM	0.15	0.79	0.51	0.25	0.79	0.56

Table 5.5: Comparison of the ML algorithms' performance when applied on different encoding techniques and external features are used. Default values in the algorithms provided from the sklearn [123], apart from the use of linear kernel for the SVC.

5.3.3 Hybrid results

The last set of experiments includes the execution of the hybrid solution which is the combination of ML algorithms followed by revision from the rule-based component. The performance of the ML algorithms is impressive when estimated with micro calculation, but also misleading due to the highly imbalanced dataset. Therefore, the rule-based component aims at increasing the F1-score in the binary calculation. Similarly to the ML-based approach, the algorithms are executed with different encoding mechanisms and there are execution batches that include the additional features. The hybrid solution is expected to have higher recall compared to ML-based solutions due to the addition of positive instances which are recognized from the rule-based component while the precision of the algorithms is expected to decrease.

Table 5.6 presents the f1-score with binary, micro, and macro calculation for the hybrid solution. In terms of f1-binary score, every hybrid solution outperforms its corresponding plain ML implementation. The performance of the proposed hybrid solution surpasses the ML solution, independently from the algorithms that have been deployed while offering a minimum standard for every deployed algorithm, which is at least equal to or surpasses the performance of the ML solution when estimated with binary calculation. The best performance is achieved when the MLP is deployed using TF-IDF uni-gram encoding reaching 0.54 f1-binary score. On the other hand, when the performance is estimated using micro calculation, the ML algorithms outperform the hybrid solutions and when the macro average is used, the two approaches present comparable results. Micro-averaged results are a measure of effectiveness for the majority class, which in our case study is the less important class. Over-

			F1-binary	F1-micro	F1-macro	F1-binary	F1-micro	F1-macro
N-gram				Hybrid			Hybrid+	
Uni-gram	TF	MLP	0.52	0.73	0.67	0.53	0.72	0.67
		DT	0.48	0.69	0.63	0.46	0.69	0.62
		LR	0.47	0.71	0.64	0.48	0.72	0.64
		SVM	0.50	0.71	0.65	0.51	0.71	0.65
	TFIDF	MLP	0.54	0.73	0.68	0.53	0.72	0.67
		DT	0.46	0.69	0.62	0.44	0.67	0.60
		LR	0.40	0.70	0.60	0.42	0.70	0.61
		SVM	0.46	0.72	0.63	0.46	0.71	0.63
Bi-gram	TF	MLP	0.52	0.74	0.67	0.48	0.73	0.65
		DT	0.48	0.69	0.63	0.47	0.70	0.63
		LR	0.46	0.72	0.63	0.47	0.72	0.64
		SVM	0.51	0.73	0.66	0.51	0.73	0.66
	TFIDF	MLP	0.52	0.73	0.67	0.50	0.73	0.66
		DT	0.45	0.68	0.61	0.44	0.68	0.61
		LR	0.41	0.70	0.61	0.40	0.70	0.60
		SVM	0.44	0.71	0.63	0.45	0.71	0.63

Table 5.6: Comparison of the hybrid solutions' performance when applied on different encoding techniques and external features are used.

all, there is a marginal deviation between the alternatives that have been deployed, around 0.48, 0.70 and 0.60 in binary, micro and macro calculation respectively.

Overall, the results present an impressive consistency indicating the significant impact of the rule-based mechanism on the hybrid solution while the impact of external features on the ML algorithms was reduced for the hybrid solution. The impact of the hybrid approach is evident when different encoding methods are compared since neither the encoding technique nor the number of tokens that are included in the encoding process has a significant impact on the algorithm's performance. Moreover, the results of the algorithms indicate the impact of the hybrid solution which increases the performance of -almost- every solution, but it eliminates the unique behaviour of each algorithm. Therefore, it is important to delve deeper into the behaviour of the hybrid solution compared to the ML algorithms.

Table 5.7 presents the difference in the performance of the two algorithms, with the hybrid solution outperforming the ML one in terms of recall but coming short in precision. For example, the MLP executed with unigram TF without any features achieves 0.61 precision and 0.36, whereas when it is integrated into the hybrid architecture it presents 0.43 precision and 0.67 recall. The highest precision is achieved with the implementation of the SVM with bi-gram TF encoding reaching 0.74, while the recalls falls to 0.30.

5.4 Discussion

The proposed hybrid solution combines the benefits of the data-driven solutions using ML algorithms while integrating domain knowledge through a rule-based mechanism. In this regard, the proposed method can identify implicit arguments that cannot be detected by the ML algorithms, a very important ability when imbalanced datasets are used due to

			Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
			ML solution				Hybrid solution			
N-gram			Encoding		Encoding + features		Encoding		Encoding + features	
Unigram	TF	MLP	0.61	0.36	0.57	0.45	0.43	0.67	0.43	0.70
		DT	0.45	0.28	0.38	0.22	0.38	0.62	0.38	0.59
		LR	0.61	0.27	0.66	0.29	0.40	0.58	0.41	0.58
		SVM	0.53	0.37	0.53	0.42	0.41	0.64	0.40	0.67
	TF-IDF	MLP	0.60	0.38	0.55	0.39	0.44	0.70	0.43	0.70
		DT	0.41	0.25	0.34	0.22	0.37	0.58	0.35	0.58
		LR	0.00	0.00	0.62	0.12	0.36	0.45	0.37	0.48
		SVM	0.71	0.15	0.57	0.20	0.40	0.55	0.40	0.56
Bi-gram	TF	MLP	0.70	0.30	0.78	0.16	0.44	0.65	0.42	0.58
		DT	0.41	0.24	0.38	0.20	0.38	0.62	0.39	0.59
		LR	0.68	0.17	0.70	0.23	0.40	0.54	0.40	0.56
		SVM	0.74	0.30	0.64	0.32	0.43	0.62	0.42	0.64
	TF-IDF	MLP	0.61	0.33	0.69	0.29	0.43	0.67	0.42	0.62
		DT	0.44	0.30	0.37	0.22	0.36	0.58	0.36	0.57
		LR	0.00	0.00	0.50	0.08	0.37	0.45	0.36	0.45
		SVM	0.69	0.08	0.60	0.16	0.39	0.52	0.39	0.54

Table 5.7: Comparison table between ML-based solution and hybrid solution. Comparison on binary performance on precision and recall.

the tendency of the ML algorithms to favour the dataset's majority class. Additionally, the concept of semantic similarity for the task of argument detection is assessed providing domain knowledge through a limited AB. However, the manual construction of the AB poses two major challenges: the rise of scalability issues because of limited coverage and the risk of bias because the AB is constructed manually after the examination of the Twitter dataset.

Moreover, a series of experiments are executed through different combinations of text encoding methods and algorithms while also using different evaluation metrics to gain a complete overview of the proposed solution. Two critical concerns were raised by the deployment of the hybrid solution: the drop in the solution's performance when evaluated with micro evaluation and the strong impact of the rule-based mechanisms that obscure the unique characteristics of the ML algorithms. Based on the experimental results, the following insights can be compiled from the experiments:

- Imbalanced data express real-world scenarios and often require a special approach, thus the use of different estimation methods (i.e. binary/micro/macro calculation) for the metrics that are used are of crucial importance for having a holistic view of the problem.
- The results from the execution of the rule-based approach indicate that simple methods, such as the Jaccard index for estimating the semantic similarity, cannot be applied to complex tasks like argument detection.
- For the creation of a balanced rule-based mechanism, attention should also be paid to the negative class because otherwise, the precision of the solution will be reduced more than the expected threshold.
- Vectorization techniques are of major importance for the argument detection task since

their effect is significantly stronger compared to the value that is offered from the additional features.

- Hybrid solutions present better results in imbalanced data due to their capability of identifying instances in the minority class, which is typically the most important one in real-world applications.

Despite the challenges that have been raised, the performance of the hybrid solution is more than promising while domain knowledge can reveal knowledge aspects in the minority class. The creation of the rule-based mechanism offers a calibration process that can enhance the precision or the recall of the hybrid solution depending on the task at hand, and thus offer tailored solutions based on the nature of the problem. The proposed hybrid solution enhances the trust and the explainability of the ML predictions, paving the way towards more explainable AI.

The three different software engineering approaches that have been applied did not present any major deviations from the projected performance, hence they could be suggested for both different NLP tasks and different use cases. More specifically, the rule-based methodology produced results with high precision and low recall because the knowledge base was designed to identify the argumentative sentences in a negatively imbalanced dataset. In different use cases with more balanced datasets, the rule-based approach is expected to produce better results. However, the AB that has been created was designed for the Nord Stream 2 dataset, hence it cannot be re-used in a different context. On the other hand, the data-driven approach of the ML algorithms can be applied to different tasks and different environments without major modifications. The result trends were the anticipated ones with the unigram and TF encoding usually outperforming the more complicated encoding methods. An aspect that should be considered is the impact of the Twitter metadata, a series of features that cannot be extracted from a different source of text. The inherent tendency of the ML algorithms to favour the majority class has been addressed using the hybrid solution, a solution that was designed to increase the recall of the ML algorithm while sacrificing its precision. The hybrid solution produced the anticipated results indicating its suitability for imbalanced datasets. Overall, the three different approaches are easily replicable without any pitfalls, except for the design and construction of the AB which is, by definition, domain-specific.

Recalling the initial research questions that have been expressed at the beginning of the thesis, this chapter covers some significant aspects that are worth researching. More specifically:

- It provides an insight into the challenges that are raised when a theoretical argumentation framework is deployed on social media statements (research question 2).

- It presents the limitations of the ML algorithms when they are deployed in real-life settings. The imbalanced nature of the Nord Stream 2 dataset poses some significant limitations in their performance (research question 3).

- It applies different methodologies for the task of argumentation detection in data derived from Twitter, revealing the impact and the social implications of argumentation detection (research question 5).

Finally, this chapter covers three different contributions, as stated at the beginning of the chapter. It demonstrates a proof-of-concept implementation for the AFAD and compares its performance with established rule-based alternative methods such as sentiment dictionary and Jaccard similarity. The results indicate the additional benefits of adding background knowledge related to the topic that is examined (contribution 3). Additionally, in this chapter, four different ML algorithms are deployed and compared with and without the use of external features. Because the Nord Stream 2 dataset was introduced in this thesis, there could not be any comparisons with previously published work, thus this goal is partially covered (contribution 4). Finally, this chapter assists in the achievement of the thesis' fifth contribution by implementing a hybrid solution through the integration of domain knowledge into ML algorithms. The results indicate the benefits of this method for the task of argumentation detection, although further testing is required to confirm these benefits in different settings (contribution 5).

The following chapter, *chapter 6 Argumentation detection in political data* presents the second use case of the thesis, applying the proposed methodology in the 2019 UK presidential debate. The proposed methodology is assessed in a different context and its performance confirms the added value of hybrid methodology for the task of argumentation detection. It is worth noting the imbalanced datasets for both use cases reveal the real-life settings that have been selected, setting additional challenges to traditional ML methods.

Chapter 6

Argumentation detection in political data

The importance of the argumentation detection task in AM pipelines becomes evident when it is applied to datasets that represent real-life scenarios with short text statements, such as a social media debate. However, social media text is not the exclusive source for short text. Transcripts from live discourse on television, radio, or podcast offer a previously unexplored alternative with high impact. The previous chapter covered the task of argumentation detection in social media; this chapter uses the transcripts from the two-hour debate aired by Sky News on April 2, 2015, having statements from David Cameron, Nick Clegg, and Ed Miliband. In the provided dataset, the non-argumentative class outnumbers the argumentative one representing real-life settings, following the same pattern that was observed in the social media dataset. The argumentative clauses are not the dominant form of expression even in a political debate.

The rest of this chapter is organised as follows. In section 6.1 an introduction to argumentation detection in political data takes place, section 6.2 presents the different methodologies that have been deployed in the political debate dataset, and section 6.3 presents their performance. Finally, section 6.4 provides an overview of the chapter's findings and initiates the discussion on the impact of the hybrid methodology and the potential of argumentation detection in political data.

6.1 Introduction

Argumentation holds its roots in rhetoric and philosophy, thus different argumentation techniques are tested in the field of political science by researching political speeches and debates. The transcripts from these interactions have the potential to offer datasets that can be used for different AM-related tasks extracting useful information such as frequency of arguments, analysis of the distinctive sub-components, and reasoning evaluation. In the meantime, the trend in using shorter sentences to express ourselves is not limited to social media platforms, but it also appears in more formal events such as political speeches [166, 118]. For example, in political debates, it is common to anticipate specific claims because the agenda and the stances of the politicians are more or less known. Therefore, even if the claims are incomplete and they do not form a valid argument, they are instantly recognised.

The main reason behind the effortless and instant grasp of the underlying argument behind a statement is the capability of the human mind to perceive the context of the statement using aspects and implications of already known information. In the process of human reasoning, the interpretation of natural language is an instant operation that combines linguistic, lexical, and sentiment characteristics of the language with the background knowledge that is previously obtained and processed. One of the most remarkable processes of the human mind is the adaptation of the background knowledge on the subject of the discussion, the reliability of the source, and, eventually, the completeness of the argument itself. The impact of the argument itself is mitigated by external factors that should be considered in the modelling of the task.

Detecting argumentation in short-text that is not acquired from online sources, but from live debates, is a topic that deserves more spotlight from the research community. It provides a new testing field where frameworks and methods can be applied to assess their suitability. Additionally, due to the rise of user-generated content on the Web, there is an increase in the heterogeneity of textual sources that could include argumentative aspects. Therefore, it is important to evaluate novel argumentation schemes and frameworks on different datasets and sources. The second case study of this thesis uses transcripts of a political debate offering an alternative data source for argumentation in short text.

In the effort of using as much as information available, different types of data, such as audio, should be considered in AM tasks covering aspects that are not covered in text-only resources. Audio resources are rarely used in NLP tasks, however, the rise and impact of podcasts could change the current status [71, 73, 26]. Especially for AM related tasks, the integration of audio into an AM pipeline takes place only, at least to my knowledge, in the work of Lippi and Torroni [95] covering a political debate including speeches from three UK presidential candidates from the 2018 election race.

On a similar note, the use of a rule-based framework in the field of AM, such as the AFAD, is not popular due to the dominance of the data-driven approaches that usually use only textual information, and in some cases, the available metadata from the online platforms that offer these data. Even if the use of these alternative sources of data does not necessarily outperform the established ML algorithms, they often provide useful insight into the nature of the problem that could be used in different contexts or use cases.

The inspiration for the creation of the AFAD was driven by the need to design an agile framework that can be deployed on short text, regardless of its original source; whether it is social media text, casual discourse, or a political debate. In the previous chapter, the proposed framework was deployed on social media text offering solid theoretical foundations, while hybrid architecture enhanced the trust and the explainability in the predictions of the ML algorithms.

This chapter explores the applicability of the AFAD in the domain of the political debate and evaluates the performance of different ML algorithms, and, eventually, tests the hybrid methodology that is introduced in the previous chapter. In summary, the main contributions of this chapter are:

- Demonstrates a second proof-of-concept implementation for the AFAD.
- Explores the alternative of using data from political speech to deploy research on argumentation detection in short text.

- Assesses the value of integrating rule-based mechanisms into ML algorithms on balanced and imbalanced datasets.
- Presents the performance of four different ML algorithms and the effect of applying additional features to them.

The rest of the chapter is organized as follows: Section 2 presents the experiments that took place in four different datasets (one dataset for every presidential candidate) and it includes rule-based solutions, ML algorithms, and the proposed hybrid architecture. Section 3 presents the results of the experiments. Section 4 initiates a discussion and presents the challenges.

6.2 Methodology

Argumentation detection in transcripts from presidential debates offers unique opportunities to extend the field of NLP domain in domains that have not yet collected a vast amount of research. The inclusion of specific features that exist in verbal communication could offer a new perspective to existing methodologies. However, in this thesis, the audio characteristics of speech are discarded, and the focus is given on the inclusion of background knowledge. Known arguments that are anticipated from each candidate are manually collected and used to implement the AFAD, while a hybrid architecture is also implemented aiming to increase the explainability of the NLP models.

In this section, the three different approaches that have been implemented are presented and compared. The rule-based mechanism, subsection 6.2.1, offers an implementation of the AFAD, subsection 6.2.2 presents the implementation of different ML algorithms, and, finally, subsection 6.2.3 proposes a hybrid solution for argumentation detection enhancing the trust in the outcome of the ML algorithms.

6.2.1 Rule-based mechanism

The objective of the rule-based approach is to assess the conceptual proximity between two chunks of text. For example, the claim *now it's key that we keep a strong economy in order to fund a strong NHS* and the political pillar of *importance of public healthcare* are correlated, even though they do not share any common words. In the use case of the political debate, the construction of the Claim DB and the Reason DB, claims and reasons are manually extracted (domain-dependent) after an evaluation process. The collected arguments highlight the main objectives of each candidate in the political debate. In the original transcripts dataset, the speech is divided into self-contained statements which are annotated as claimed / non claimed. After a careful evaluation, it has been noticed that the claimed statements are supported by a valid reason even if the reason is not given in the same statement or the same dataset. The political positions and reasoning processes of each political party are already known.

Figure 6.1 illustrates seven examples of political statements in the dataset, presenting three complete arguments having claims and reasoning, two arguments having a claim supported by implicit reasoning, and two statements without any claim. The first two examples are taken from Cameron's speech, the first one presents a claim (*"keep a strong economy"*)

A01 now it's key that we keep a strong economy in order to fund a strong NHS
A02 and the plan is working because last year we had the fastest growing economy of any of the major Western countries
A03 I don't equally think it is fair to do what the Labour Party wants to do, which is actually to increase borrowing: that doesn't help the future generations
C04 I think it's a dismal choice, cutting too much or borrowing too much
C05 I say: "Britain could do so much better than it's done over the last 5 years."
NC06 but David you just said that you were tackling tax avoidance
NC07 you know, I got this sort of pious stance from Ed Miliband

Figure 6.1: Arguments in different means, in different topics sharing a common word

providing the reason underneath the claim (*"fund a strong a NHS"*) while the second argument expresses a claim on the economy plan (*"the plan is working"*) supported by evidence (*"we had the fastest growing economy of any of the major Western countries"*). The third example is from Clegg's speech expressing a claim (*"increase borrowing"*) and supporting it with a reason (*"doesn't help future generations"*) stressing the need for sustainable policies. The fourth example includes a claim (*"it's a dismal choice, cutting too much or borrowing too much"*) taken from Clegg's speech without any explicit reasoning but justified with common sense. In the fifth statement, Miliband expresses a claim (*"Britain could do so much better than it's done over the last 5 years"*) supported by a series of statements later in his speech. The first three of the aforementioned examples are self-contained expressing valid arguments, while others, although valid, require an amount of background knowledge. On the other hand, the sixth statement (*"but David you just said that you were tackling tax avoidance"*) from Milliband's speech, does not include any claim and it criticizes a political opponent. Similarly, the last statement (*"you know, I got this sort of pious stance from Ed Miliband"*) from Clegg's dataset is another criticism of the political opponent without including any claim.

The examples from the political debate, even though short, differ from the statements that are usually found on Twitter. They do not have any typos, there is no use of slang, and, overall, they include less noise despite the use of shortened verbal forms. Therefore, the proposed theoretical frameworks need to provide the necessary flexibility to cover different real-life settings. The AFAD provides a great degree of adaptability allowing the adjustment of the mapping functions depending on the context and the priorities of every use case. For example, in this use case where the dataset is transcripts of a live political debate, it is expected to meet a vocabulary with extensive use of words that trigger the audience's feelings while it is also expected to find arguments that support the political agenda of every candidate. Therefore, the speech of every candidate was studied and a list of words that represent their political stances was created. Political stances have a strong connection with argumentation since they often express core values of different political beliefs, hence claims and reasons often overlap.

The construction of the AB for the political debate dataset followed a slightly different approach compared to what was described in the previous chapter. The statements for the construction of the Nord Stream 2 gas pipeline extracted from Twitter cover a plethora of

topics and even though they tend to follow the latest news, they cannot have the direct interaction as that in a live political debate. In the later use case, the context, the set-up, the background of the speakers, and even the tone of their voices, have a major effect on the evaluation of the expressed statements. In a live political debate it is not expected from the participants to proceed in a deep political analysis, but rather trigger emotions and memories from the audience on political stances that are well-known and construct the major beliefs of each party.

Example	**Claim(s)**
what my plan is about is basically one word: security. security for you, for you family, for our country	security as priority, family importance
and I want to see the NHS move to much more seven-day operation like your GP being open 8 in the morning 8 in the evening all the way through the week	extensive working schedule
I don't equally think it is fair to do what the Labour Party wants to do, which is actually to increase borrowing: that doesn't help the future generations	balanced economy, future generations
so I guess my approach to immigration could be summarised simply as this: that I want Britain to be open for business but not open to abuse	immigration policy
I think it's much better to have a fair plan which says those with the broader shoulders should bear the greatest burden, and we will make reductions in spending	fair taxing system, spending reductions

Table 6.1: Examples of argumentative statements and their claims that are used to construct the AB of the political dataset.

Table 6.1 provides examples of statements and their expressed claims that have been used in the political debate by its participants. From the five examples that are presented in the table, only one provides a complete justification *I don't equally think it is fair to do what the Labour Party wants to do, which is actually to increase borrowing: that doesn't help the future generations*, while the other examples rely on previous knowledge. The aforementioned example introduces a specific claim which is against increased borrowing and reasons that this policy will not assist future generations. In line with the findings of the previous chapter, background knowledge plays a crucial role in the evaluation of the statements as argumentative or not, and it contrasts the common belief of expecting political leaders to justify their claims. Similar to the approach that has been followed in the previous chapter, the claims have been transformed into uni-grams for easier manipulation.

The lack of providing reasons in the live political debate is justified by two main aspects: the limited time provided for the participants to express their positions, and the wide audience they refer to. These two restrictions force the candidates to stick to known and well-communicated political stances of the parties they participate in, hence the justification of each claim seems redundant within the pace of the discussion. For example, it is expected from the conservative party to promote the social value of the family and to prioritise public security. The validity of the arguments is not evaluated, and it is not in the scope of this thesis which focuses on argumentation detection in short text in real-life settings.

Table 6.2 presents the words that imply the existence of an argumentative segment in the transcripts. The first row illustrates the words Cameron uses to express arguments and create the AB for this use case; some of them indicate core political values such as *family, national, security* while others indicate political claims on specific topics such as

the Brexit referendum. In Clegg's case the words that indicate argumentative segments can be grouped into clusters based on the topic they face - economic plan (*borrowing, cutting, reduce*), youth population (*youngsters, kids, generation*), and migration policy (*immigration, freedom, claim*) - providing a solid ground to create the AB. Finally, in Mibiband's dataset, there is a lack of coherence in the keywords that are used to express arguments which are mostly descriptive, also exhibiting the longer dataset compared to the other two candidates. In contrast with the previous case study, the construction of the AB did not split into sub-components due to the implicit reasoning that follows the majority of the claims. For instance, a stance supporting the public healthcare system can be easily justified. For this case study, the AB was created based on the analysis of the transcripts, alternative methods could create the AB solely based on the known political stances of every party since the AFAD does not pose any limitations.

Candidate	Keywords
Cameron	family, national, security, referendum, growing economy, union, week, improve, important
Clegg	borrowing, cutting, reduce, needs, european union, youngsters, kids, generation, population, mental health, immigration, freedom, claim
Miliband	turn round, create, gp, opportunity, succeed, guarantee, apprenticeships, generation, concerns, disaster, unqualified, teachers, priorities, false

Table 6.2: List of keywords that indicate the existence of argumentation in each political transcript.

Similarly to the use case in detecting argumentation in social media text, two different encoding techniques have been used: TF and TF-IDF. After the encoding process, the estimation of the cosine similarity takes place, determining the semantic distance between the content of the AB and the statements in the transcripts of every political candidate. This method has been proved to yield beneficial results when social media datasets are studied and it is expected to produce similar results when applied in the context of a political debate.

6.2.2 ML algorithms

The constantly increasing use of ML algorithms for a series of computational problems, including various NLP tasks, has allowed non-expert users to exploit state-of-the-art techniques and algorithms using simple interfaces that are offered by the ML libraries. ML algorithms allow researchers to focus on the development of data processing pipelines and produce decent results instead of struggling to develop algorithms for every use case. Additionally, computer scientists and AI experts have extended their field of work including new areas such as bioinformatics and linguistics without requiring a deep theoretical background, offering a new perspective to these domains. For example, in the use case of the political debate in this chapter, there is no need for a political analysis from an expert but only the extraction of textual features.

In this case study a wide set of features has been used to assist the performance of the ML algorithms including 1) lexical features, uni-gram techniques with TF and TF-IDF have been deployed, 2) semantic features, the NLTK PoS tagger has been used which can identify

and group words into different categories that display similar syntactic behaviour, and 3) sentiment features, the TextBlob software has been used determining the degree of polarity and subjectivity.

The ML algorithms that have been selected are multi-layer perceptron (MLP), decision tree (DT), logistic regression (LR), and support vector machines (SVM).For consistency reasons and ease of comparisons, the ML algorithms have been executed with the same parameters that were used in the Nord Stream 2 dataset. More specifically, the MLP is implemented using one hidden layer with 100 units while the rectified linear unit (ReLU) has been deployed in the hidden layers and the logistic sigmoid function in the output layer presenting complexity of $O(n * m * h^k * o * i)$ where n is the training samples, m the features, k the hidden layers, h the neurons, and o the output neurons. In the political debate dataset, the execution time of the algorithm is significantly shorter, because the training samples are significantly fewer and there are not any Twitter metadata that could be used as features. The DT has been implemented without maximum depth, and the Gini criterion has been used to evaluate the quality of the split; the minimum number of samples to split the internal node is 2 and for the creation of a leaf it is 1; all the available features have been used to decide the best split, the number of leaf nodes has not been defined, the two classes have been assigned with the same weight, no minimum impurity value has been defined, and the cost complexity pruning value has been set to 0. The LR has been implemented using the limited-memory BFGS (L-BFGS) solver allowing up to 100 iterations before convergence, the L2 penalty has been set for regularization, the inverse of the regularization strength is set to 1.0, the primal formulation has been selected and the same weight has been assigned to the two classes. Finally, for the SVM algorithm, the linear kernel has been selected for its implementation, the squared l2 penalty has been set to 1, the shrinking heuristic has been enabled, the tolerance for stopping criterion is set to 1e-3, the cache size has been set to 200 MB, the two classes have been assigned to have the same weight, and finally, no limit on iterations within the solver has been set. More technical details regarding feature engineering and the process of selecting algorithms are provided in the previous chapter.

Finally, the same pre-process function that has been deployed in the rule-based component has also been applied here, stripping the text from unnecessary noise, although the noise is expected to be significantly reduced since the transcripts are manually compiled and any filled pauses or filler words are not recorded. Additionally, there is no use of hashtags or mentions in the transcripts. Another significant aspect that is different from the social media use case is that there are three distinctive datasets, hence it is expected **that** the ML algorithms do not have consistent performance in every dataset. Moreover, taking into consideration the tendency of the ML algorithms to favour the majority class, the Miliband dataset is expected to present the worst results in terms of F1-binary score since the non-argumentative class is 69.2% of the dataset. On the other hand, the ML algorithms on the other two datasets are expected to have more predictable performance.

6.2.3 Hybrid approach

A known weakness of the ML algorithms is their tendency to favour the majority class which, in real-life settings, it is often not the research's point of interest, creating multiple false negatives. On the other hand, the rule-based algorithms are designed to identify positive instances even in harsh conditions, thus they often falsely label positive instances (false positives). Tak-

ing this behaviour into consideration, for this case study the hybrid architecture, presented in the previous chapter, is deployed. In the proposed architecture the predictions of the ML algorithms are fed into the rule-based component which is responsible to fix any inconsistencies that are identified.

An alternative to the proposed hybrid pipeline is the integration of the extracted keywords as features for the ML algorithms. In this alternative approach, there are two options, every keyword acts as a binary feature, or the occurrence of the keywords is added, producing a numerical value. For the use case of the political debate, this alternative is rejected, after preliminary analysis. A different, and more complicated, implementation of the hybrid architecture could include different weights for the keywords in the list. However, this approach was rejected because it would require a political science expert without ensuring better performance. The proposed hybrid methodology balances the uncontrolled data-driven approach and the logic programming paradigm. Therefore, it is expected to increase the f1-binary score in every dataset, and in the more balanced datasets (Cameron, Clegg) an increase in the f1-macro score is also expected.

If the hybrid architecture outperforms the data-driven solution in the use case of the political debate, it indicates that it has the potential to be applied to different datasets for the task of argumentation detection after the necessary updates in the creation of the AB. On the other hand, the design process of the proposed solution presents a disadvantage compared to data-driven solutions; it requires domain knowledge. Since the creation of the AB is a manual process, it hinders the expansion of the method to different datasets even if they have similarities.

6.3 Results

The evaluation of the algorithm's performance is achieved using three different metrics (precision, recall, F1-score), each one presenting a different aspect of the algorithms' behaviour. The F1-score has been calculated using two different techniques, binary and macro calculation, aiming at covering any doubts that may arise due to the imbalanced nature of the data. On imbalanced datasets reporting only the F1-score it is often not enough and it could be misleading. Therefore, the estimation of the metric is achieved using binary and macro calculation which calculates metrics for each label and finds their unweighted mean without taking into consideration the label imbalance. When the minority class is valued the most, the macro calculation is the preferred way of calculation because it treats both classes as equal regardless of the number of instances in each class.

The rest of the section is organised as follows. Subsection 6.3.1 presents the performance of the rule-based approach, subsection 6.3.2 reports the results from the ML algorithms, and subsection 6.3.3 presents the results of the hybrid solution.

6.3.1 Rule-based approach

For the first set of experiments, the rule-based approach was implemented with different implementations of AFAD for the functions $g(r_i, t)$ and $h(c_j, t)$ representing the mapping from reasons to claims and claims to arguments respectively. The first scoring function, $sem\_AB(x)$ function, unifies the g(r,t) and h(c,t) functions since, for the creation of the AB

in this case study, claims and reasons are not split into different sub-databases. For the text encoding process, the TF and TFIDF techniques have been used. The sub(x) function has also been used, evaluating the subjectivity of the input statements with the use of an external dictionary [165]. Finally, for the last approach, a combination of the $sem\_AB(x)$ and the $sub(x)$ functions was implemented ($comb(x)$ function), where a statement has to display both high semantic similarity with the known claims in the AB using the $sem(AB)$, while it also reaches a relatively high subjectivity score assigned from the $sub(x)$. Two more methods were also used as benchmarks; the first benchmark is a random function and the second one is the Jaccard similarity coefficient score. In the rule-based approach, statements that have a semantic similarity score over the median are labelled as argumentative offering a universal implementation without considering the distribution between argumentative and non-argumentative statements in the specific datasets.

	Models	Cameron				Clegg				Miliband			
		Prec	Rec	F1	F1*	Prec	Rec	F1	F1*	Prec	Rec	F1	F1*
Rule-based	Baseline	0.51	0.54	0.53	0.54	0.40	0.46	0.43	0.46	0.32	0.53	0.40	0.50
	Jaccard Sim.	0.52	0.56	0.54	0.56	0.62	0.70	0.65	0.67	0.36	0.59	0.45	0.54
	sem_AB(x)	0.64	0.68	0.66	0.67	0.69	0.78	0.73	0.75	0.42	0.69	0.53	0.60
	sem_AB(x)_idf	0.64	0.68	0.66	0.67	0.65	0.74	0.69	0.71	0.32	0.53	0.40	0.50
	sub(x)	0.49	0.53	0.51	0.52	0.60	0.67	0.63	0.65	0.31	0.51	0.39	0.49
	comb(x)	0.65	0.39	0.48	0.59	0.77	0.52	0.62	0.70	0.42	0.35	0.38	0.57

Table 6.3: Comparison of the different approaches and algorithms that have been used in three different datasets. F1* is F1-score macro calculated

Table 6.3 presents the results produced after deploying rule-based algorithms to three different datasets, reporting the performance in terms of precision, recall, binary f1-score, and macro f1-score. From the aforementioned metrics, more attention is given to the macro f1-score because it does not favour the majority class. The $sem\_AB(x)$ function surpasses the rest of the methods in every one of the three datasets since the dictionary that has been developed is tailored for identifying even implicit arguments reaching 0.67, 0.75, and 0.60 macro F1-score in Cameron, Clegg, and Miliband datasets, respectively. The $sem\_AB(x)\_idf$ presents results equal to sem_AB(x) only in the Cameron dataset, while in the other two datasets its performance is significantly lower, reaching 0.71 and 0.50 f1-score macro in the Clegg and Miliband datasets, respectively. The combination of the two approaches ($comb(x)$ function) presents lower but comparable results to $sem\_AB(x)$ boosting the precision, but lowering the recall, and eventually reaching 0.59, 0.70, and 0.57 in the Cameron, Clegg, and Miliband datasets, respectively. The Jaccard similarity index presents better results to the random baseline, but significantly worse to the best-performer function of $sem\_AB(x)$. Finally, the general-purpose lexicon ($sub(x)$ function) underperforms in Cameron and Miliband datasets while only in Clegg datasets it outscores the random baseline achieving 0.65 f1-score macro.

6.3.2 ML-based approach

Table 6.4 presents the performance of the ML algorithms when executed having as input n-grams with and without additional features. The first group presents the performance of the algorithms having as input solely 1-gram, and the second group the performance when additional features are also included. In the Cameron dataset, the LR and SVM present the best performance both achieving 0.66 F1-score macro; in the Clegg dataset the MLP with external features outperforms the competition with 0.71 F1-score macro, and in the Miliband dataset, the SVM reaches 0.70 F1-score macro. The algorithms present a similar pattern in the Cameron and Miliband dataset; the SVM and the LR present the highest scores -the SVM reaches 0.66 f1-score macro in the Cameron and 0.70 f1-score macro in Miliband, and the LR reaches 0.66 f1-score macro in Cameron and 0.69 f1-score macro in Miliband- while the MLP comes third (0.60 F1-score macro in the Cameron, 0.66 F1-score macro in the Miliband), and the DT underperforms in those two datasets (0.61 F1-score macro in the Cameron, 0.60 F1-score macro in the Miliband). On the other hand, in the Clegg dataset, the MLP and the DT present the highest scores reaching 0.70 and 0.68 f1-score macro, respectively, while the SVM falls to 0.57 f1-score macro. Overall, the use of external features has a positive impact only in the Clegg dataset, with the MLP reaching 0.71 F1-score macro while in the other two datasets the impact is either negative or insignificant.

	Models	Cameron				Clegg				Miliband			
		Prec	Rec	F1	F1*	Prec	Rec	F1	F1*	Prec	Rec	F1	F1*
ML	MLP	0.57	0.60	0.58	0.60	0.67	0.65	0.66	0.70	0.60	0.43	0.50	0.66
	DT	0.62	0.49	0.55	0.61	0.68	0.57	0.62	0.68	0.46	0.43	0.44	0.60
	LR	0.70	0.53	0.60	0.66	0.64	0.39	0.49	0.60	0.74	0.41	0.53	0.69
	SVM	0.65	0.60	0.62	0.66	0.59	0.37	0.45	0.57	0.64	0.51	0.57	0.70
ML+	MLP	0.57	0.60	0.58	0.60	0.71	0.63	0.67	0.71	1.00	0.04	0.08	0.45
	DT	0.45	0.40	0.43	0.49	0.53	0.50	0.52	0.58	0.46	0.39	0.42	0.60
	LR	0.62	0.60	0.61	0.64	0.68	0.61	0.64	0.69	0.70	0.39	0.50	0.67
	SVM	0.60	0.60	0.60	0.62	0.65	0.57	0.60	0.66	0.50	0.47	0.48	0.63

Table 6.4: Comparison of the different approaches and algorithms that have been used in three different datasets. F1* is F1-score macro calculated

Compared to the rule-based methods, the ML algorithms present higher precision and lower recall indicating that a hybrid method could improve the recall by enhancing the performance of the solution. Additionally, a hybrid method could enhance the trust and explainability in the outcome of the NLP solutions, increasing the confidence in the outcomes of the AI solutions.

6.3.3 Hybrid results

Table 6.5 presents the performance of the hybrid solutions including the audio-enhanced ML, rules-enhanced ML algorithms (hybrid), and rules-enhanced ML algorithms with the use of external features (hybrid+). In the first group of results, the three datasets have been tested on a hybrid approach deploying Google Speech API and ground truth transcripts

enhanced with audio features [95]. For the learning system, the SVM has been deployed using original terms, part-of-speech tags, and lemmas, while for the audio features the RastaMat library is used to extract statistical information for the mel-frequency cepstrum coefficients (MFC). Audio features improved the performance of the algorithms in the Cameron and Clegg dataset reaching 0.61 and 0.69 f1-score macro, whereas in the Miliband dataset they failed to improve the performance due to higher noise in the original audio snippet. In the original paper, the results are reported only on f1-score macro, hence the comparison is limited only to one metric, which is considered the preferred method in imbalanced datasets due to its ability not to favour the majority class. In the proposed hybrid approach, the integration of the sem_AB(x) into the ML algorithms improves the binary F1-score while its impact on the macro F1-score is negligible. More specifically, in the Cameron dataset the hybrid component improves the f1-score by 5% the DT and LR algorithms, and by 6% the MLP and SVM algorithms. In the Clegg dataset, the improvement in the performance of the algorithms is significantly higher, with the hybrid component increasing the f1-score in the MLP 10%, in the DT 13%, in LR 20%, and the SVM 24%. Finally, in the Miliband dataset, in terms of f1-score, the MLP is improved by 4%, the DT and LR by 5%, and the SVM decreases by 1%. In terms of f1-score macro, the algorithms show a significant improvement only in the Clegg dataset with the MLP improving by 6%, the DT by 8%, the SVM by 12%, and the 14%. Similar to what has been noted in the execution of the ML algorithms, the inclusion of additional features does not seem to improve their performance.

	Models	Cameron				Clegg				Miliband			
		Prec	Rec	F1	F1*	Prec	Rec	F1	F1*	Prec	Rec	F1	F1*
Audio	GroundTruth	—	—	—	0.55	—	—	—	0.51	—	—	—	0.59
	GroundTruth + Audio	—	—	—	0.61	—	—	—	0.59	—	—	—	0.63
	GoogleSpeech	—	—	—	0.48	—	—	—	0.51	—	—	—	0.31
	GoogleSpeech + Audio	—	—	—	0.53	—	—	—	0.53	—	—	—	0.29
Hybrid	MLP hybrid	0.56	0.75	0.64	0.60	0.68	0.87	0.76	0.76	0.48	0.63	0.54	0.65
	DT hybrid	0.59	0.61	0.60	0.62	0.69	0.83	0.75	0.76	0.41	0.59	0.49	0.59
	LR hybrid	0.63	0.67	0.65	0.66	0.68	0.70	0.69	0.72	0.53	0.63	0.58	0.68
	SVM hybrid	0.63	0.74	0.68	0.67	0.66	0.72	0.69	0.71	0.49	0.65	0.56	0.66
Hybrid+	MLP hybrid	0.57	0.75	0.65	0.61	0.70	0.85	0.76	0.77	0.45	0.39	0.42	0.59
	DT hybrid	0.51	0.60	0.55	0.54	0.57	0.67	0.62	0.63	0.44	0.61	0.51	0.62
	LR hybrid	0.60	0.74	0.66	0.65	0.67	0.78	0.72	0.73	0.52	0.63	0.57	0.67
	SVM hybrid	0.58	0.74	0.65	0.62	0.66	0.80	0.73	0.73	0.43	0.65	0.52	0.61

Table 6.5: Comparison of the different approaches and algorithms that have been used in three different datasets. F1* is F1-score macro calculated

6.4 Discussion

The extensive experiments provide some useful insight into both the value and the adaptability of the AFAD. It also presents interesting results regarding the performance of the proposed hybrid method for the task of argument detection increasing the macro-F1 score, indicating its suitability for imbalanced datasets where the detection of the minority class is

of crucial importance.

The audio-enhanced ML solutions also offer a boost in the performance of the algorithms when the GroundTruth transcripts are used, but their performance does not surpass the proposed hybrid methodology that relies on the concept of AB. Additionally, from both hybrid approaches, it is clear that argumentation detection is a challenging task that requires manual intervention, whether it is GroundTruth transcripts or the creation of an AB. On the other hand, the use of the external features on the n-gram approaches does not seem to offer valuable information. However, in different scenarios such as a debate in social media where the additional features carry valuable information through metadata, it is suggested to be included in the experiments.

The use of audio signals is not extensively used in the field of AM, while the fusion of different types of features, such as audio, textual, and visual for more popular NLP tasks shows some significant potential [130, 134]. On a similar note, the concept of the knowledge base has not been extensively utilised for AM-related tasks. Combining different types of data could pave the way for novel approaches to the wider field of NLP including aspects previously ignored. Especially, the design of a framework that provides the foundation for future expansions including different types of data could have a significant impact on showing adaptability in different environments because it includes different types of features.

The AFAD shows great potential since it provides a great starting point for the deployment of different methods and the combination of different approaches. The theoretical foundation of modelling assists in both understanding the task and focusing on aspects that could have a strong impact on the argumentation detection task. Additionally, the implementation of the AFAD can enhance the performance of the ML algorithms. Regarding the performance of the proposed hybrid solution, despite the challenges that have been raised, its performance is promising since domain knowledge can reveal knowledge aspects in the minority class. Additionally, it offers a calibration process that can enhance the precision or the recall of the hybrid solution depending on the task at hand, thus tailored solutions based on the nature of the problem can be proposed.

Similarly to the observations described in Section 5.4, the hybrid method has been found to achieve the best results, although the use of the claim base is domain-specific and can limit its generalisation to different domains. More specifically, the rule-based methodology produced results with high precision and low recall because the knowledge base was designed to identify the argumentative sentences in a negatively imbalanced dataset. In different use cases with more balanced datasets, the rule-based approach is expected to produce better results. However, the AB that has been created was designed for the Nord Stream 2 dataset, hence it cannot be re-used in a different context. On the other hand, the data-driven approach of the ML algorithms can be applied to different tasks and different environments without major modifications. The result trends were the anticipated ones with the unigram and TF encoding usually outperforming the more complicated encoding methods. An aspect that should be considered is the impact of the Twitter metadata, a series of features that cannot be extracted from a different source of text. The inherent tendency of the ML algorithms to favour the majority class has been addressed using the hybrid solution, a solution that was designed to increase the recall of the ML algorithm while sacrificing its precision. The hybrid solution produced the anticipated results indicating its suitability for imbalanced datasets. Overall, the three different approaches are easily replicable without any pitfalls, except for

the design and construction of the AB which is, by definition, domain-specific.

Overall, in this case study, a series of experiments were executed trying different combinations of text encoding methods and algorithms while also using different evaluation metrics gaining a clear view of the potential of the AFAD. The implementation of the hybrid methodology provides an insight for the exploitation of the background knowledge for the task of argumentation detection. Based on the experimental results, a synopsis of the results is presented:

- The AFAD has been deployed on one more dataset that represents real-life settings providing comparable results with state-of-the-art ML algorithms.
- The use of more than one evaluation metric is of major importance on imbalanced datasets since one metric can be misleading for the performance of an algorithm.
- The results from the execution of the rule-based approach confirm the results from the social media use case that the Jaccard index is a simplistic method that cannot be applied for the task of argumentation detection.
- Hybrid solutions, in the majority of the cases, enhance the performance of the ML algorithms to their capability to identify instances in the minority class, which is typically the more important one in real-world applications.

Argumentation techniques have changed over time and short text has impacted the way people express themselves. As a result, the argumentation analysis of political debates can take place using schemes and frameworks designed for short text. The use of symbolic AI in hybrid architecture has significant benefits because specific arguments from presidential candidates are anticipated. Additionally, the predictions of the ML algorithms can be adjusted, hence more explainable solutions can be produced. Finally, even though the environment is more controlled compared to social media text, it still represents real-life settings with non-argumentative statements outnumbering the argumentative ones.

Recalling the initial research questions that have been expressed at the beginning of the thesis, this chapter covers some significant aspects that are worth researching. More specifically:

- It uses a previously published dataset illustrating the relevance and the importance of the task of argumentation detection (research question 1).
- It questions the dominance of the data-driven approaches through the performance of the hybrid solution in the Clegg dataset (research question 3).
- It is the second case study on argumentation detection, applied to political data, showing the wide spectrum of applications for the specific task (research question 5).

Finally, this chapter covers three different contributions, as stated at the beginning of the chapter. It provides an implementation for the AFAD (different from the implementation showcased in the previous chapter) and compares its results with trivial rule-based approaches. The results confirmed the findings from the first use case; tailored background knowledge offers a significant boost on the task (contribution 3). Additionally, four different

ML algorithms have been tested providing an overview of data-driven approaches. Their performance has also been compared with results from previously published research and outperformed them (contribution 4). Finally, the fifth contribution is also answered in this chapter. A hybrid approach is implemented by integrating domain knowledge into ML algorithms increasing the explainability of the results of the algorithms while also boosting the performance in terms of binary f1-score. In combination with the findings in the previous chapter, it confirms the added value of a hybrid architecture when deployed in real-life settings (contribution 5).

The following chapter, *chapter 7 Transfer Knowledge* researches the problem of transferring knowledge in different domains. First, an introduction to the term contextual embeddings takes place, then the BERT architecture is illustrated and implemented, and details regarding its implementation are provided. Finally, the hybrid architecture is implemented on top of the BERT. The results indicate that contextual embeddings provide a viable alternative, which, however, require careful fine-tuning, while the integration of the rules enhances the trust in the performance of the AI solution.

Chapter 7

Transfer knowledge

The complex nature of natural language, the adoption of informal language, and the influence of social media on shaping society's opinions indicate the need for introducing advanced NLP methods capable of providing a deeper understanding of the natural language. These changes pose challenges to traditional argumentation schemes due to their lack of adaptability while data-driven solutions, such as deep learning, seemed to have reached their limits due to hardware limitations and the constantly increasing request for new data. This condition caused a shift of AI models towards general intelligence either by transferring knowledge between different tasks or by integrating symbolic AI that allows them to reason more similarly to the way humans do. The previous chapters showcased the impact of background knowledge in two different case studies, on social media text, and debate transcripts. This chapter explores the capability of existing language models to transfer knowledge from different NLP tasks into argumentation detection when deployed in real-life settings.

The rest of the chapter is organized as follows: section 7.1 provides an introduction to the need for transferring knowledge between different NLP tasks and the emergence of contextual embeddings. Section 7.2 describes the methodology that is followed providing both the theoretical background on contextual embeddings and the practical details on the implementation of the BERT architecture. The results from the BERT implementation and the hybrid solution are given in section 7.3, and, finally, in section 7.4 a discussion is initiated on the challenges that have appeared, the need for transferring knowledge between different tasks, and the potential of the BERT architecture.

7.1 Introduction

The field of NLP has experienced significant growth in the last decade due to the progress in ML and the development of related commercial applications. There is a need to develop techniques that can exploit the volume of text that is generated from Web-based applications such as social networks, microblogging, social review sites, community blogs, etc. The field of NLP includes a series of distinctive tasks with different goals on multiple communication levels. On a basic communication level, there are tasks such as word segmentation, parsing, and part-of-speech tagging, while there are also higher-level NLP tasks such as dialogue management and summarization that require an abstract understanding of the human language. The field of AM is the evolution of computation argumentation while it is also influenced

by the task of opinion mining. An AM pipeline usually integrates aspects such as argument and stance detection. The produced results could be useful in different scenarios, like legal reasoning or public legislation, especially on controversial topics. It is important to adopt a cross-task approach in the NLP and try to transfer knowledge between different tasks and domains, especially in emerging areas with limited annotated datasets.

The first attempts at modelling argumentation using AI created the field of computational argumentation where a set of rules describe the fundamental mechanisms humans use in argumentation. The modelling allows the interpretation of AI algorithms. Even though data-driven solutions have dominated the NLP research, there are still rule-based mechanisms that are widely used, such as sentiment dictionaries, and grammar structure techniques, while there are some very successful commercial solutions (WordNet, ChatScript, etc.).

Rule-based solutions rely on the concept that some background knowledge is necessary to interpret and model natural language, a notion close to what is called common sense. The rule-based algorithms aim to find the correlation between a known set of arguments and a statement where a higher correlation indicates closer semantic proximity. On the other hand, data-driven approaches use ML algorithms to reveal information that cannot easily be extracted through human effort. In addition, there is a demand for developing research methods that enhance the trust and the explainability of the AI services by integrating symbolic AI. These changes pave the way towards argumentation detection in short text, exploiting methods that seem to have the potential to transfer knowledge between tasks and domains, yet not leaving the outcome of these methods completely uncontrolled.

The recent advances in the domain allow the use of off-the-shelf algorithms from ML libraries that can be deployed without requiring expert knowledge. Except for the encoding of the human language, ML algorithms can also utilize external features such as syntax trees, social media metadata or audio extraction, but the additional features often bring more noise than useful information. The majority of the studies follow a data-driven approach, which, although successful and easy to use, does not promote trust and explainability of the AI methods. The third alternative is the hybrid approach that combines the rule and ML-based methods into a pipeline to utilize the benefits each one presents. The hybrid pipelines usually adjust the ML predictions based on the domain knowledge while they are capable of adjusting their sensitivity based on the task at hand and the needs of the environment.

From the latest advances in the wider domain of NLP, the concept of contextual embeddings has emerged. Based on the encoder/decoder architecture, the contextual embeddings acquire information for multiple tokens (e.g., words in British novels) and represent the information in an encoded data structure with per-token outputs. In contrast with traditional embeddings, contextual embeddings examine the context of the token and adjust the encoded vector. The integration of context could offer a significant boost in every AM-related task since human reasoning is based to a great extent on context gained from previous knowledge.

The previous chapters illustrated that the task of argumentation detection in short text can be assisted with the use of a manually created knowledge base. If the pre-trained embeddings can replace -to some extent- the manual creation of the knowledge base it would save time and resources. Rule-based, ML, and hybrid methods have been deployed on four different datasets providing insight into the benefits of the hybrid methodology, especially in imbalanced datasets. In this chapter, the concept of contextual embeddings is studied and implemented, discovering viable alternatives for transferring knowledge between different NLP

tasks.

Similar to argumentation detection, the majority of the NLP tasks such as opinion mining, sentiment analysis, and source evaluation require annotated datasets, a process that includes the design of an annotation strategy and the training of the participants. Overall, an intensive and expensive process. Additionally, although the tasks are semantically close, they answer different questions, thus the same dataset cannot easily be used in different tasks. Therefore, unless there are no limitations on the resources, there is a need to try novel methods capable of transferring knowledge between tasks.

BERT is considered one of the most novel approaches in the wider field of NLP with multiple applications in different fields and tasks, including the field of AM. However, due to the inherited complexity of the argumentation process, the field of application is still focused on legal text and argumentative essays. In previous research work the BERT model has been applied to legal text on practical case law from the European Court of Human Rights for three different tasks: argument clause recognition, argument relation mining, and argument component classification, presenting better performance than ELMo and GloVe in most of the argument mining tasks [192]. Legal text presents significant differences compared to the datasets that are used in this thesis that contain short statements with often implicit reasoning.

Apart from legal text, another popular source of input for AM-related tasks is persuasive essays. The application of BERT for detection of argumentative relations using the full paragraph context to the model provides better results for the specific task [50], while on a more extensive comparison the BERT model demonstrates comparable performance to previous state-of-the-art methods in two different datasets [189]. However, similar to legal text, persuasive essays provide a more solid structure compared to social media datasets. A middle ground between a Twitter discussion and a persuasive essay is provided within the subreddit *change my view (CMV)* which has been used for inter-component relation prediction stressing the importance of the context that is used to train the BERT model [167].

Identifying the need for adopting methods capable of transferring knowledge to new domains, the BERT architecture is expanded and modified to evaluate the performance of contextual embeddings for the task of argumentation detection in different datasets. Then, there is a comparison between the produced results and with a series of techniques including rule-based approaches, ML algorithms, and hybrid methods that have been introduced, implemented and evaluated in the previous chapters.

Overall, contextual embeddings are tested in four datasets, from two different topics, and two different sources providing a holistic overview of their performance. In summary, the main contributions of this chapter are:

- The use of the contextual embeddings, more specifically the BERT architecture, for the classification problem of argumentation detection in short text.
- The use of alternative input vectors in the BERT model focuses on the nature of short text by prioritizing the length of the statement and ignoring the position of the tokens.
- The presentation of extensive comparisons between different AI methods on multiple datasets for the task of argumentation detection allows us the comparison of different techniques.

The rest of the chapter is organized as follows: Section 2 presents a technical introduction to CE, the fundamental principles for the BERT architecture, and details regarding its implementation. Section 3 presents the results of the BERT architecture and the implemented hybrid architecture. Section 4 initiates a discussion and presents the challenges.

7.2 Methodology

The methodology that has been followed in this chapter is differentiated from what is provided in the last two case studies. The first case study provides a proof-of-concept implementation of the AFAD, deploys four different ML algorithms, and tests a hybrid solution. The methodology in the second case study presents a similar pattern while its results confirm the added value of adding background knowledge into the ML algorithms, enhancing both the performance and the explainability. This chapter attempts to exploit a generic language model to extract valuable information.

The following subsections provide a complete overview of the use of contextual embeddings in the argumentation detection task. In the first subsection, an introduction to contextual embeddings takes place, revealing the motivation behind the development of the BERT architecture and its use for the task of argumentation detection. The second subsection provides technical details on the BERT architecture. Finally, the third subsection presents the results of the fine-tuned architecture in four different datasets.

7.2.1 Contextual embeddings

Natural language is a sequence of characters with limited utility as they cannot be fed directly to computational models, hence a series of characters with variable lengths has limited utility compared to numerical features. With the use of embeddings, words are expressed as real-valued vectors in a pre-defined vector space and they are developed based on the usage of the words on the input text during the training process. The outcome of the learning process produces similar vectors for similar words, naturally grasping their meaning while capturing syntactic and semantic regularities.

In traditional word embedding methods, like word2vec [108] or GloVe [129], words are always mapped to the same vector, failing to identify the different meanings a word might have based on a specific context. Therefore, their use in a cross-domain environment is questionable. NLP systems need to transfer knowledge expanding their application fields to new domains. The argumentation techniques in different topics and different means differ significantly, however, modern argumentation detection systems must develop a deep semantic understanding of different topics in different contexts.

Figure 7.1 visualizes the different meanings the word *cold* can get under different circumstances creating different vectors. Three examples are provided illustrating the different uses of the word, while every instance of the word is represented with a different vector creating a highly contextual representation. The traditional embeddings create a unique vector for each word failing to encode additional information and eventually grasp the wider context around the word. Contextual word embeddings such as ELMo [131] and BERT [33] offer a solution because they take into consideration the context of every word providing a better representation depending on the words that follow and precede. The contextual embeddings

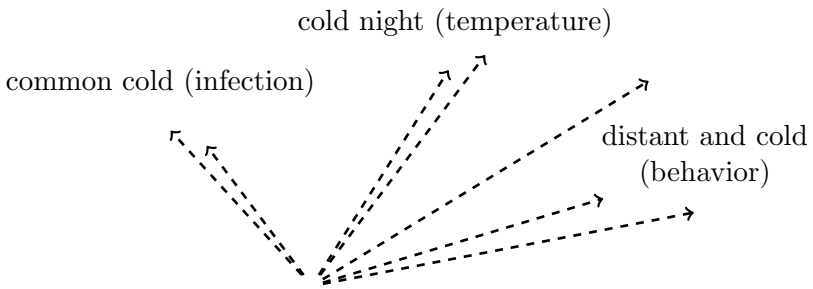

Figure 7.1: A graphical illustration of the contextual embeddings.

are produced from large, pre-trained encoders trained in extensive datasets, often in more than one tasks, such as textual entailment and sequence tagging. Contextual embeddings are often used for downstream tasks through the addition of layer(s) adjusting the knowledge of the system towards specific tasks and domains.

7.2.2 From RNN to BERT

Recurrent neural networks such as the Long Short-Term Memory (LSTM) network are designed for sequential processing. In sequential processing the input sentences are processed word by word and the knowledge in the network is maintained through past hidden states. The encoding of a specific word is retained only for the next step, meaning that the representation of a word affects the representation of the following one, hence its influence is slowly vanishing. As a result, the representation of a word in a sentence is affected only by the words that are preceded in close proximity while the impact of the words that are further is minimized and the words that are followed are ignored. Argumentation is a complex process that often requires compiling information that is not located in close proximity or is stored in a common knowledge repository. Therefore, there are some limitations in the performance of the RNN in the task of argumentation detection.

The evolution of the RNN is the encoder-decoder architecture, a two-component scheme where the first part, the encoder, reads and maps the input sequence into a fixed-length vector. The second component of the architecture, the decoder, interprets and maps the vector back to a target sequence [25]. The traditional format of the encoder-decoder model receives a sequence of words and forms another sequence of words as an output. The weakness of this architecture is the bottleneck that is created due to its sequential nature while engaging solely the most recent layer. The method of attention allows the model to look further than the last position in the input sequence for characteristics that allow it to better understand the input sequence and eventually assist in a better encoding process [64]. The attention has been used with great success in the field of machine translation enhancing its performance on the sentence level. In the context of argumentation detection, it could allow us to identify implicit argumentation schemes. A novel architecture that uses the encoder/decoder scheme without the deployment of a RNN is the Transformer [177] which presents impressive results in deep learning tasks such as machine translation.

BERT (Bidirectional Encoder Representations) is a unique Transformer-based language model using only stacked encoders outperforming existing solutions such as ELMo [131],

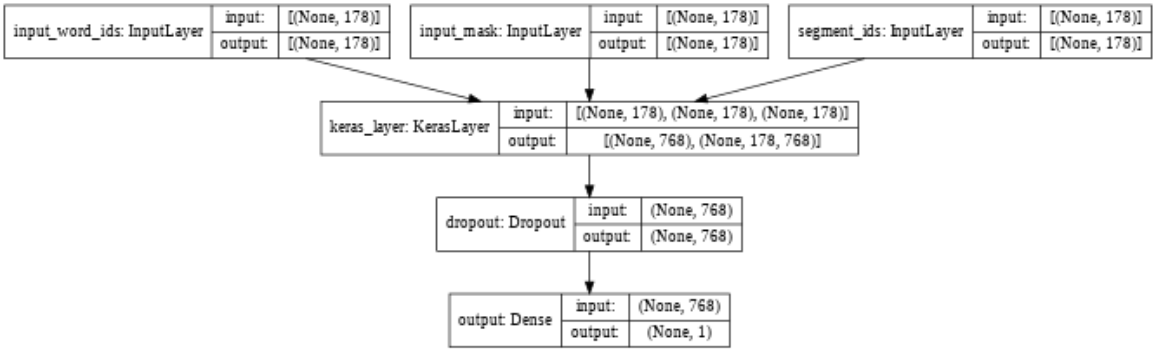

Figure 7.2: The architecture of the BERT for the task of argumentation detection

ULM-Fit [63], and Open AI Transformer [135] in numerous NLP and NLU tasks. A key aspect of BERT's good performance in different tasks is the use of semi-supervised learning, thus the model is trained to understand the patterns of the language. BERT is designed on the basis of carrying out two different objectives, Masked Language Model (MLM) which is also known as the fill in the blanks task, and Next Sentence Prediction (NSP), hence it is suitable to be deployed on different NLP tasks simultaneously. A second breakthrough for BERT is the use of bi-directional context. The model takes into consideration the words that both precede and follow, allowing it to gain a much better understanding of the context in which the word(s) was used. The original release of BERT includes two models: 1) $\text{BERT}_{\text{BASE}}$ having 12 Encoders with 12 bi-directional self-attention heads, and 2) the $\text{BERT}_{\text{LARGE}}$ having 24 Encoders with 24 bidirectional self-attention heads. Both of them are trained from unlabeled data using as sources the English Wikipedia and the BooksCorpus with 2,500M and 800M words, respectively.

The application of the BERT architecture on Twitter data for the task of AM-related tasks has not been extensively tested, and, at least to my knowledge, it has not been applied for the task of argumentation detection in online short text. Additionally, BERT is trained using sentences from formal resources, hence it is unclear whether it will be able to adapt to the additional challenge of short sentences. The deployment of BERT in different datasets from different resources on an emerging NLP task could reveal previously unexplored capabilities of the contextual embeddings in real-life settings.

7.2.3 Implementation of BERT

Since the datasets that are used are relatively small, and there is an overall scarcity of annotated datasets for the tasks of argumentation detection, BERT architecture can assist by adjusting the downstream task. BERT's multi-task training settings enable its use for both sequence-based tasks (pair of sentences) such as QA and classification tasks, and binary classification tasks, such as sentiment analysis, with the former following the findings of the MLM and the latter of the NSP. In AM related tasks, the closer tasks to MLM and NSP are argument component classification and relation prediction, but their benefits in argumentation detection are not yet proven. This novel use of BERT is examined on the level of understanding of the human language while it also examines its effectiveness on different sources of input.

Figure 7.2 illustrates the BERT architecture that has been designed and fine-tuned for the task of argumentation detection in this thesis. The first layer depicts the three 178-dimensional input vectors representing: 1) the input tokens encoded through BERT embeddings, 2) a binary mask indicating the length of the incoming chunk, and 3) a second binary mask identifying the BERT reserved character for the start of every sentence (segment embedding). Compared to the inputs suggested in the original BERT paper, the positional embeddings have been replaced by the length input mask. The $\text{BERT}_{\text{BASE}}$ pre-trained model is in the second layer receiving three different input vectors and producing pooled and sequence vectors as outputs. The pooled vector pools the 178 dimensions of the sequence vector into one dimension and it is used as an input to the dropout layer with a 0.5 rate to prevent overfitting. Finally, in the last layer of the architecture, there is a dense layer using the sigmoid activation function producing the outcome of the architecture. In the last layer, four different variations of the Adam optimizer have been tested: Adam, Nadam, Adamax, and Adam Weight Decay, while each one of them has been tested with 3 different learning rates: 2e-4, 2e-5, and 2e-6. The different learning rates provide a relatively wide area for experiments while offering an extensive overview of the performance of the BERT architecture. The Nord Stream 2 and Miliband dataset is expected to present some additional challenges due to the dominance of the negative class. Local optima are expected to be found, leading to overtraining despite the additional dropout layer which is expected to resist this change. Moreover, on highly imbalanced datasets, there is the risk of classifying all the instances as positives.

The size of the pre-trained BERT model and the required resources for its execution are well-known problems, which are analogous to the size of the dataset. The training phase of the BERT model on a new domain requires significant resources that are often not easily accessible, hence the majority of the work in the domain focuses on fine-tuning the pre-trained model for the downstream task. Additionally, the available implementations of BERT were limited at the time of writing this thesis, thus the TensorFlow platform was used via the Keras APIs that are similar to the scikit-learn's, the "gold standard" of modern machine learning APIs and it has been used for the deployment of the traditional ML algorithms that have been presented earlier in the thesis. The problem of the required resources for the execution of the experiments was partially bypassed using Google Colab, an online code editor that executes the code not on the local machine, but on the cloud. The deployment set-up, including libraries, interfaces, and the code editor, is developed by Google, hence minimum interoperability issues are expected to appear.

7.3 Results

The two previous chapters implemented three different approaches to tackle the problem of argumentation detection in short text in real-life settings. This chapter examines the ability of a language model to transfer knowledge between different NLP tasks. Subsection 7.3.1 presents the performance of the BERT model in four different datasets using four different activation functions with different learning rates, and subsection 7.3.2 presents the performance of the hybrid solution when the rule-based component is placed on top of the BERT model.

7.3.1 BERT performance

In this scenario, the evaluation of the algorithms' performance takes place calculating the F1-score using three different estimation techniques, binary, micro, and macro calculation. It is important to detect any weak aspects in the BERT architecture for the task of argumentation detection, hence it is important to calculate the F1-score with three different techniques. Especially on heavily imbalanced datasets, reporting only one metric could be misleading for the performance of the algorithms.

	F1*	F1**	F1**	F1*	F1**	F1**	F1*	F1**	F1**	F1*	F1**	F1**
	Nord Stream			Cameron			Clegg			Miliband		
Adam (2e-4)	0	0.78	0.44	0.55	0.38	0.28	0.68	0.63	0.61	0	0.67	0.40
Adam (2e-5)	0.55	0.80	0.71	0.62	0.73	0.70	0.32	0.59	0.51	0	0.67	0.40
Adam (2e-6)	0	0.78	0.44	0.41	0.54	0.52	0.67	0.69	0.69	0.33	0.67	0.56
Adamax (2e-4)	0	0.78	0.44	0	0.62	0.38	0.69	0.75	0.74	0	0.67	0.4
Adamax (2e-5)	0.24	0.79	0.56	0.42	0.62	0.57	0.65	0.63	0.62	0	0.67	0.4
Adamax (2e-6)	0	0.78	0.44	0.41	0.46	0.46	0.62	0.69	0.68	0.44	0.63	0.58
Nadam (2e-4)	0	0.78	0.44	0	0.62	0.62	0.5	0.69	0.64	0	0.67	0.4
Nadam (2e-5)	0	0.78	0.71	0.35	0.70	0.58	0.69	0.66	0.65	0.38	0.73	0.60
Nadam (2e-6)	0	0.78	0.44	0.43	0.57	0.54	0.60	0.63	0.62	0.00	0.67	0.40
AdamWeightDecay (2e-4)	0.35	0.79	0.61	0.00	0.62	0.38	0.00	0.56	0.36	0.71	0.75	0.75
AdamWeightDecay (2e-5)	0.31	0.80	0.60	0.38	0.68	0.57	0.65	0.63	0.62	0.11	0.67	0.45
AdamWeightDecay (2e-6)	0	0.78	0.44	0.42	0.49	0.48	0.38	0.59	0.54	0.52	0.50	0.50

Table 7.1: Results of the neural network architecture taking advantage of BERT in terms of F1 binary, F1 micro and F1 macro. Four different algorithms have been deployed with 3 different learning rates.

Table 7.1 presents the results of the implemented models deployed on 4 different datasets while calculating the F1-score with three different methods (binary, micro, macro). In the Nord Stream 2 dataset due to the highly imbalanced nature of the source, the algorithm often finds local optima without identifying true positive samples in the dataset, hence the binary f1-score is often zero. The Adam optimizer with learning rate 2e-05 presents the best results in this dataset demonstrating high scores on every metric, 0.55, 0.80, and 0.71 in f1 binary, f1-micro, and f1-macro respectively, outperforming any other solution. The problem of local optima also appears, to a lesser extent, on the Cameron dataset when the learning rate is the highest one in three different optimizers. The algorithm with the best performance uses Adam optimizer with a learning rate of 2e-05 favouring the majority class achieving f1 binary score of 0.62, f1 micro 0.73, and f1 macro 0.70. The f1 binary score is relatively low, while the performance in terms of f1 micro and macro calculation is comparable to the hybrid approach presented in the previous section. On the Clegg dataset, the implementation using Adamax optimizer with a 2e-04 learning rate achieves 0.69 f1 binary score, 0.75 micro score, and 0.74 macro score, presenting comparable results with the hybrid methodology. Its performance on the f1-binary score is relatively low, indicating a tendency to favour the majority class. Finally, the problem of local optima is evident in the Miliband dataset as 6 different implementations do not identify true positives following a similar pattern to the Nord Stream 2 dataset. The Adam Weight Decay implementation with a 2e-04 learning rate presents the best results achieving a 0.71 f1-binary score, 0.75 f1-micro score, and 0.75 f1-macro score.

Overall, the performance of the BERT presents great fluctuation raising questions about its trust, especially on highly imbalanced datasets. Similar to traditional ML algorithms, the BERT architecture favours the majority class, which, often, is the least important one.

The similar pattern that is observed between the traditional ML algorithms and the implementation of BERT illustrates the common principles between the two approaches. The different datasets that have been used indicate that the BERT model can be generalised relatively well if the fine-tuning process proceeds the deployment. However, its performance in limited and heavily imbalanced datasets creates some concerns about its application in every domain and task. The addition of background knowledge into a hybrid architecture could stabilise the fluctuation in the performance and enhance the aspects of explainability and robustness.

7.3.2 Hybrid on top of BERT

The last two chapters demonstrated the potential of the hybrid architectures by improving the performance of traditional ML algorithms in terms of binary F1-score. Integrating knowledge into hybrid architecture indicates a viable alternative for the task of argumentation detection. Therefore, the rule-based mechanism as described in the previous section was deployed on top of the BERT architecture compiling BERT's predictions with the arguments stored in the AB, aiming to increase the detection of positive instances. More specifically, the BERT and the rule-based predictions are fed into the hybrid decision mechanism which prioritises the detection of more positive instances, even if it means decreased performance in the majority class. In BERT implementations where no positive instances were identified, the hybrid predictions are identical to the rule-based methodology, indicative of the BERT's poor performance.

	F1	F1*	F1**	F1	F1*	F1**	F1	F1*	F1**	F1	F1*	F1**
	Nord Stream 2			Cameron			Clegg			Miliband		
Adam (2e-4)	0.47	0.62	0.59	0.55	0.38	0.27	0.61	0.44	0.30	0.45	0.54	0.53
Adam (2e-5)	0.48	0.62	0.59	0.68	0.65	0.64	0.75	0.75	0.75	0.45	0.54	0.53
Adam (2e-6)	0.47	0.62	0.59	0.59	0.54	0.54	0.82	0.81	0.81	0.53	0.56	0.56
Adamax (2e-4)	0.47	0.62	0.59	0.63	0.68	0.67	0.72	0.69	0.68	0.45	0.54	0.53
Adamax (2e-5)	0.47	0.62	0.59	0.63	0.62	0.62	0.88	0.88	0.88	0.45	0.54	0.53
Adamax (2e-6)	0.47	0.62	0.59	0.55	0.46	0.44	0.79	0.78	0.78	0.58	0.58	0.58
Nadam (2e-4)	0.47	0.62	0.59	0.55	0.38	0.27	0.80	0.81	0.81	0.45	0.54	0.53
Nadam (2e-5)	0.49	0.63	0.60	0.57	0.46	0.43	0.81	0.81	0.81	0.45	0.54	0.53
Nadam (2e-6)	0.47	0.62	0.59	0.65	0.68	0.67	0.76	0.75	0.75	0.45	0.54	0.53
AdamWeightDecay (2e-4)	0.47	0.62	0.59	0.55	0.38	0.27	0.80	0.81	0.81	0.49	0.56	0.55
AdamWeightDecay (2e-5)	0.44	0.57	0.55	0.56	0.54	0.54	0.85	0.84	0.84	0.45	0.54	0.53
AdamWeightDecay (2e-6)	0.47	0.62	0.59	0.57	0.51	0.50	0.77	0.78	0.78	0.54	0.46	0.44

Table 7.2: Hybrid on top of BERT architecture for the best implementation on each dataset.

Table 7.2 presents the results of the BERT hybrid methodology providing a complete overview of the method. In the Nord Stream 2 dataset, the integration of the rule-based mechanism increases the performance in 11 out of 12 implementations. However, it fails to improve the performance of the best implementation while the influence of the rule-based mechanism is so strong that erases any characteristic for the different implementations producing 0.47 f1 binary score, 0.62 f1 micro score, and 0.51 f1 macro score -or a similar score-

for every implementation. On the Cameron dataset, it improves the f1-binary score for every implementation, it has a negligible impact on f1-micro and f1-macro in the majority of the implementations, but it lowers the performance in the best implementation with Adam optimizer and learning rate 2e-05 presenting f1 binary score 0.68, f1 micro 0.65, and f1-macro 0.64. On the Clegg dataset, the hybrid methodology presents impressive results since it increases the performance in terms of f1-binary in every implementation, and there are only two implementations that do not also improve in terms of f1 micro and f1 macro. The best instance, Adamax with a 2e-05 learning rate, achieves a 0.88 f1 score. In the Miliband dataset, there is an increase in the performance of all the implementations, except when using the previously highest performance implementation of Adamax optimizer with a 2e-4 learning rate.

Overall, the hybrid solution offers an alternative, increasing the majority of the implementations, but in some cases, it might not be beneficial to the best existing solutions. The hybrid methodology is a trade-off between trust in the AI solution and peak performance. The implementation of a hybrid architecture enhances the replicability of the solution dealing with the problem of the local minima. The few outliers that could produce the highest performance are sacrificed for the sake of a higher level of explainability and transparency. However, in different commercial applications, it is more important to offer a safety net to the algorithm's predictions rather than aiming at the outlier that will produce the best results.

7.4 Discussion

The adjustment of ML predictions through topic-specific rules in a hybrid architecture is a method that declines in popularity in contrast to data-driven solutions. However, hybrid solutions present an alternative solution increasing, in the majority of the instances, the f1 binary score. In this regard, hybrid solutions can identify implicit arguments that cannot be detected by the ML algorithms, a very important ability when imbalanced datasets are used, due to the tendency of the ML algorithms to favour the dataset's majority class. Furthermore, the benefits of pre-trained contextual embeddings for the task of argumentation detection have been stressed through the use of BERT presenting comparable results to domain-specific methodologies. If we take into consideration that the contextual embeddings can initially be trained in shorter chunks of text, alternative implementations can yield even better results. For example, a BERT model could be trained on a massive Twitter database and then fine-tuned for the different NLP tasks that would be executed for different tasks in a noisy environment.

A series of experiments were executed in four different datasets trying three different approaches while also using different evaluation metrics providing a complete overview of the implemented methodologies. Contextual embeddings have changed the map in NLP-related tasks offering off-the-shelf reliable solutions for a wide range of tasks, provided that they are correctly fed and fine-tuned. After a preliminary analysis and testing of the BERT architecture in the task of argumentation detection in short text, one of the input vectors has been modified to focus on the size of each chunk and not on the position of every token inside the chunk. This decision is merely driven by the use of natural language in social media where syntax rules are neglected and the use of hashtags does not follow any predefined order.

From the experiments in this chapter, the implementation and deployment of the BERT

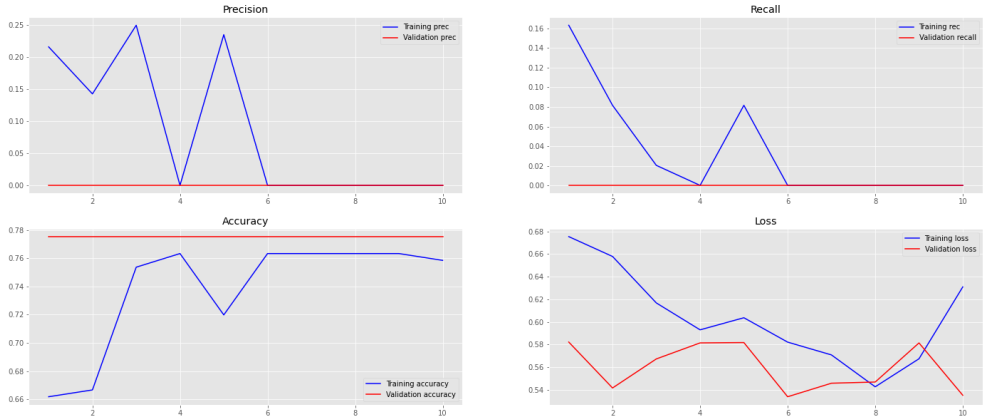

Figure 7.3: BERT deployment with Adam optimizer with learning rate=2e-5 and dropout rate=0.5

architecture, two points of discussion have emerged: a) optimal fine-tuning is not easily achievable, and b) human supervision is required to analyse the results of the experiments. The alternative of an exhaustive search for hyper-parameters has not been chosen because it carries the risk of overtraining due to the lack of human evaluation. In the provided scenarios, after multiple testing iterations, the dropout rate has been set to 0.5 to avoid the overfitting that is caused due to the small size of the dataset. Figure 7.3 depicts the flat line in precision, recall, and accuracy for the deployed architecture when the dropout rate is set to 0.1 and the learning rate is 2e-5. The values of the binary cross-entropy loss indicate an inability of the model to learn from the input text. The behaviour is quite often in the heavily imbalanced datasets, thus there were cases in which the model did not positively label any instance from these datasets. Overall, based on the experimental results, the following insights can be compiled from the implementation of the hybrid solution and the fine-tuning process of BERT:

- Imbalanced data express real-world scenarios and often require a special approach, thus the use of different estimation methods (i.e. binary/micro/macro calculation) for the metrics that are used is of crucial importance for having a holistic view of the problem.
- The more imbalanced the dataset the more likely it is for the neural network to be stuck into a local optimum. In the Nord Stream 2 dataset, there are 8 implementations where the neural network does not predict a true positive instance whereas in Miliband there are 6 different cases.
- A pre-trained model, such as BERT, if fed and fine-tuned correctly, can adjust to new domains presenting comparable results to topic-specific methods.

Another aspect that should be discussed is the difficulty of the reproducibility of the experiments. Even though fine-tuning is a process significantly lighter compared to the initial training, it is still a computation process that requires significant processing power, hence the use of GPU parallelization is required. In the deployed experiments, all the pseudo-generated random seeds that are used in algorithms have been defined, but the need for

parallel execution presents some minor changes at the hardware level. The reproducibility in the results of the BERT model in limited datasets could create some minor issues that can be overcome if the researcher decides to execute the model in a local machine, by removing any parallelization strategies in the background or by cross-validating the results via multiple iterations. Nevertheless, the difference in every iteration is not expected to be important and it can be ignored.

Overall, the standard BERT model seems to generalise despite some minor problems caused by the limited size of the datasets and the different characteristics of the language it has to encounter. The application of the hybrid architecture indicates the potential benefits of domain-specific pre-training in interdisciplinary downstream tasks, with a special language context, whether it is noisy short text or legal documents [192]. However, the lack of domain-specific training did not hinder the BERT implementation from presenting comparable performance to previous state-of-the-art methods in the cases where the learning rate had the correct value.

The main research question of this chapter is what and how much a pre-trained BERT model knows for the detection of arguments in short chunks of text, and if it can be compared to domain-specific architectures. Despite the challenges that have been raised, both hybrid solutions and contextual embeddings architectures offer promising results, with each one approaching the task of argumentation detection from a different perspective. Domain knowledge can reveal knowledge aspects in the minority class and contextual embeddings can present satisfactory results in previously unknown domains with minimum human supervision.

Recalling the initial research questions that have been expressed at the beginning of the thesis, this chapter covers some significant aspects that are worth researching. More specifically:

- It connects the task of argumentation detection with the BERT model, one of the latest advancements in the NLP domain, indicating the connection of the AM with modern research advancements (research question 1).
- The extensive experiments testing the BERT model examine the limits of a deep learning architecture, revealing its shortcomings when deployed in real-life settings (research question 3).
- The deployment of the BERT model demonstrated that it is possible for generic language models to transfer knowledge for the task of argumentation detection (research question 5).

Finally, this chapter covers two different contributions, as stated at the beginning of the chapter. First, it provides the necessary theoretical and practical background for contextual embeddings, and then examines their applicability for the task of argumentation detection, assisting towards the completion of the sixth contribution. The BERT architecture has been modified focusing on the short nature of the input text producing satisfactory results and outperforming traditional ML algorithms. However, it should be noted that the high fluctuation in the performance depending on the learning rate indicates a lack of consistency and the need to increase the trust in the outcome of the deep learning architecture. To tackle

the challenge of explainability in the performance of data-driven solutions, a hybrid architecture was deployed integrating domain knowledge, confirming the added value of domain knowledge into data-driven approaches, and supporting the fifth contribution as stated in the introduction of the thesis. The results from the hybrid approach minimised the fluctuation of the deep learning architecture, increasing the explainability of its results.

The following chapter, *chapter 8 Conclusions & Future work*, summarises this thesis and evaluates the goals that have initially been set. Every chapter offers a different perspective on the objectives that were set at the beginning of the thesis. It starts by introducing the reader to the field and offering the necessary theoretical and technical background, presenting extensive experiments on the two different case studies and exploring the alternative of contextual embeddings. Additionally, future directions are provided indicating potential research goals and extensions of this thesis. The automatic creation of the AB could provide a boost to hybrid methodologies, while the extensive comparison of semantic similarity methods could clarify the needs for the NLP tasks that required a higher level of understanding.

Chapter 8

Conclusions & Future work

The main research question that initiated and drove this thesis is how to extract valuable information from natural language in real-life settings. The constantly increasing generation of textual information via social media, and the need for quantifying qualitative aspects of human language led the research towards the field of AM. The challenges that have been raised in applying traditional AM schemes on Web-generated data have shifted the focus of the research towards the task of argumentation detection.

The dominance of data-driven architectures provided satisfactory results in the last years, but they have raised concerns about the trust of the outcome because the performance cannot get easily interpreted and explained. In this thesis, background knowledge is collected and used on a hybrid solution stressing its benefits of increasing the performance on highly imbalanced datasets while also enhancing the trust in the outcome of the solutions. To minimise the human effort in collecting knowledge, the method of contextual embeddings has been tested offering satisfactory results, but their use has raised concerns about their use in real-life settings on limited datasets.

The findings indicate that domain knowledge enhances the performance of argumentation detection while also increasing the trust in the outcome. However, the collection of the domain knowledge remained a manual process revealing some limitations on its scalability. The integration of reliable clustering techniques could minimise human interaction while providing solid domain knowledge.

The following section, *Section 8.1 Conclusion*, concludes the thesis by summarising its main findings and evaluating the goals of the thesis. The section *Section 8.2 Future goals* presents suggestions that could enhance the findings of the thesis and assist research on their work.

8.1 Conclusion

The main objective of the thesis is to extract valuable information from human language as it is expressed in real-life settings. The area of NLP includes a series of fields and tasks following different perspectives and covering different needs. From all the emerging areas, the field of AM has been selected because it has the potential to quantify qualitative aspects of human language. The wider field of AM has been researched providing detailed literature focusing on the peculiarities of short text as expressed in real-life settings, mostly through social media

platforms. The choice to work in real-life settings raised some significant technical challenges (additional noise, imbalanced datasets, etc.), however, there is room for improvement and market potential. Taking these challenges into consideration, the focus was given to the task of argumentation detection. The existence or not of argumentative elements in a statement is of crucial importance because it defines the nature of the statement while this information can further be used on different NLP tasks.

Even though the task of argumentation detection has been previously researched, there were not any formal definitions neither for the argument per se nor for its components, at least to my knowledge. The need for providing the related definitions has emerged while the value of a theoretical framework has also appeared. Therefore, seven different definitions and the AFAD were introduced covering every aspect of argumentation in short text.

The next milestone of the thesis was the creation of a dataset that represents a real-life debate. The literature review indicated the lack of available datasets and a tendency towards topics and sources that present high homogeneity. The use of Twitter as the main data source disrupted many of the previously established guidelines because the arguments were often implicit or relied on specific domain knowledge. The statements that have been collected express thoughts, opinions, and reflections regarding the debate on the construction of the Nord Stream 2 gas pipeline. The topic was not previously explored, at least to my knowledge, by other NLP studies, and it paves the way for related studies on similar topics exceeding the limits of Brexit, Grexit, and political elections.

After the creation of the dataset, different implementations of the AFAD were tested, outperforming trivial metrics such as the random baseline, and the Jaccard index. The recall in the rule-based approaches was relatively high while the precision was not at satisfactory levels. Additionally, four different ML algorithms were applied and presented slightly better results in terms of binary F1-score, high precision, and low recall. The need for combining the two methods was created. A hybrid architecture was proposed and implemented increasing the performance of the algorithms while also enhancing the aspects of trust and explainability in the algorithm's behaviour.

To ensure the suitability of the proposed hybrid solution for the task of argumentation detection, a second case study took place. The dataset included short statements, but its source was not Twitter; there were transcripts from the presidential UK 2019 elections. The collection and annotation were completed on different research without my engagement, removing any implicit bias that could have occurred in the first use case. The hybrid solution exploiting background knowledge produced better results than a previously suggested hybrid solution that includes audio features. However, the collection of domain knowledge remained a manual process, stressing the need for novel methods that can transfer knowledge between domains.

The concept of contextual embeddings has emerged in the last years providing a new perspective in the wider NLP domain. The BERT architecture, originally developed by Google, provides impressive results on different tasks and it has revolutionised the entire field. However, as with every deep learning architecture, there is some ambiguity on how the final decisions are drawn raising some trust issues. The inclusion of the BERT architecture into a hybrid solution increased the explainability of the approach, and minimised the fluctuation in the performance, but it did not improve the performance of the best model.

The following subsection *8.1.1 Summary of thesis* summarises the contributions of every

chapter of the thesis, and the subsection *8.1.2 Evaluations of thesis goals* examines to what degree the thesis' objectives have been achieved.

8.1.1 Summary of thesis

Chapter 1 introduces the reader to the subject of the thesis providing an overview of the domain and presenting the advances in AI that led to the creation of multiple sub-domains including the field of NLP. Then, by delving deeper into breaking the field into tasks, the need for further investment in tasks that provide a greater level of understatement of the human language and a reliable explainability in the performance of the AI solutions has been noted. An obstacle that has appeared is the shift in the use of language towards a more casual form, mostly due to the dominance of social media in public debates. This shift has blurred the lines between opinions, arguments, and personal reflections. In this environment, detecting argumentation on short text should be regarded as a fundamental research question and a key task for the wider NLP domain. Inspired by these changes, the motivation of the thesis is provided, enhancing the trust and the explainability of NLP models by adopting a hybrid methodology for the task of argumentation detection. Next, the aims and the objectives are clearly stated, examining different aspects of the field while setting the milestones for the thesis. A short literature review on the state-of-the-art methods in the AM followed, indicating future directions. Finally, the main contributions of the thesis are presented emphasizing the relevance and the impact of the thesis.

Chapter 2 provides the necessary theoretical background presenting the related work in the wider field of AM. The chapter presents a complete literature review of the field starting by explaining basic concepts such as opinion mining and feature engineering and concluding with an evaluation of the existing research methodologies for different AM tasks, including relations identification, stance detection, and reliability-related tasks. Special focus is given on the task of argumentation detection which is the main topic of the thesis.

Chapter 3 presents the modelling process of quantifying argumentation. The chapter suggests a quantification model for argumentation in short text providing a series of definitions to support a holistic approach to the proposed framework. The quantification of argumentation assists in the deeper understanding of qualitative aspects of the natural language, hence it is necessary to achieve explainable AI systems increasing the trust in the technology. The chapter is used as a link between the literature review and the technical chapters that follow.

Chapter 4 presents the mechanisms and implications of the annotation process. A new dataset is created and annotated from Twitter on the construction of the Nord Stream 2 gas pipeline. The dataset covers a previously unexplored topic for the task of argumentation detection which suffers from a lack of annotated dataset.

Chapter 5 presents the first case study of the thesis applying rule-based, ML-driven, and a hybrid approach implementing the AFAD. The findings of chapter 5 present the performance of the different methodologies indicating that hybrid methodologies outperform existing solutions in terms of F1-binary score. The imbalanced nature of the dataset requires special handling since the positive instances are the minority in the dataset, even though they are the most important ones. The real-life settings pose some additional challenges in the deployed AI struggling to identify the positive instance in the dataset. Therefore, the use of background knowledge offered by the AB offers the desired increase in the recall.

Chapter 6 demonstrates the second case study of the thesis was rule-based, ML-driven and

a hybrid implementation of the AFAD was deployed. The use of the second dataset provides more confidence in the findings from the first use case and confirms the added value of the proposed hybrid methodology. The dataset is extracted from the UK presidential candidates from the 2018 election race and it offers a different perspective compared to the first use case. Chapter 6 illustrates how short argumentative statements are used in a political speech and evaluates the proposed methodologies on three different datasets. The results indicate that the use of an AB yields better performance, especially in terms of F1-binary score, while the integration of audio features does not seem to outperform ML algorithms that have integrated external features.

Chapter 7 explores the trending subject of transferring knowledge among different NLP tasks. The concept of contextual embeddings is realised by fine-tuning the BERT model which is designed for machine translation and is tested on the downstream task of argumentation detection. In terms of performance, the BERT model outperforms the previous methodologies in the two heavily imbalanced datasets, the Nord Stream 2 and the Miliband datasets. Additionally, a rule-based mechanism was applied on top of the BERT model presenting an alternative to the hybrid methodology which increases the performance in the majority of the implementations. The findings provide a great overview of the different methodologies that can be applied in the field, offering an extensive comparison between methods, algorithms, and datasets. The performance of the BERT architecture for the task of argumentation detection in imbalanced datasets demonstrates the potential of contextual embeddings in real-world problems. Since quality annotated datasets are not easily available, pre-trained models present an opportunity for different downstream tasks, as long as they provide the flexibility for fine-tuning and extension.

At the end of the thesis, there is Appendix A. It presents state-of-the-art tools for general NLP tasks and tools for argument search, retrieval and annotation.

8.1.2 Evaluation of thesis goals

The main aim of this thesis, as stated in the introduction, is to enhance the trust and explainability of data-driven models for the task of argumentation detection in short text. This has a chained impact on various related tasks both in the field of AM and in the wider field of NLP. This thesis has focused on exploring alternatives that can integrate background knowledge into ML algorithms, aiming at more reliable AI solutions. This aim has been achieved by accomplishing four objectives: (1) offer the theoretical framework that enables the application of different methods allowing practical implementation while the solutions are scientifically rigorous, (2) apply different algorithmic approaches in a dataset collected from Twitter over a sociopolitical debate, (3) confirm the findings of the second sub-problem in a different dataset, in the UK presidential debate 2018 debate, (4) test the potential of transferring knowledge that is gained from different NLP tasks into the task of argumentation detection.

The first objective is resolved in Chapter 3 by providing a series of definitions in the wider context of AM and introducing the AFAD. The AFAD provides a quantification method for assessing the existence of argumentation in short text, offering solid research foundations for the task of argumentation detection. The second objective is presented in Chapter 5 where 6 different rule-based and 4 different ML algorithms are deployed, offering a detailed overview of the performance of the state-of-the-art solutions for the task-at-hand. The implementation

of the AFAD is also presented outperforming the existing solutions. The impact of the hybrid solution is confirmed in Chapter 6 where the AFAD is implemented on a different dataset demonstrating better performance compared to existing hybrid solutions that engage audio features instead of an AB. Finally, the problem of transferring knowledge between different tasks in the NLP is examined in Chapter 7 where the BERT model is fine-tuned for the task of argumentation detection offering comparable results with existing solutions.

8.2 Future goals

Methods proposed as part of this thesis can be extended in several possible ways or can be generally used in other applications. Argumentation detection can be used as a preliminary task on any AM-related task while it can also provide useful insight for qualitative aspects of the dataset in every NLP task. Regarding the hybrid methodology that has been proposed, the integration of stored knowledge can be beneficial for different NLP tasks, enhancing the trust in the predictions of the ML methods and the trust towards future AI systems. However, there are two main obstacles to the wider adoption of the thesis' findings. First, the creation of the AB is not automated and it requires a time-consuming manual process, hence there are some obstacles in its application to new topics. The latest findings in the domain of the NLP stress the need for transferring knowledge to new domains and tasks. Therefore, there is a need for a method that could collect the basic arguments with minimum human supervision. The second obstacle was the construction of a reliable metric that can identify the semantic similarity between two chunks of text.

The different encoding methods involve different rationales, hence their performance could be different when calculated with different methods. Especially for the task of argumentation that requires a deeper understanding of the human language, alternative methods for estimating semantic similarity could provide better and more robust results. Therefore, an in-depth overview of semantic similarity estimation methods should take place specifically for the task of argumentation detection. Based on the aforementioned challenges, two future directions that could provide significant knowledge are automated clustering and the research for alternative methods for estimating the similarity.

8.2.1 Automatic clustering

The manual process of the AB construction could be enhanced with the use of automatic argument discovery approaches that are capable of identifying arguments in previously unseen text. The task of argument discovery presents many similarities to topic modelling, thus unsupervised algorithms will be integrated into the existing AM pipeline. The construction of a database of claims is primarily a manual process, hence the capabilities of transferring knowledge across domains are limited. In order to encounter this limitation, the manual process could be enhanced with the use of automatic claim discovery approaches that are capable **of identifying** claims in previously unseen text. The task of argument discovery presents many similarities to topic modelling, thus unsupervised algorithms will be integrated into the existing AM pipeline. The task of claim discovery presents many similarities to topic modelling, therefore the Latent Dirichlet Allocation (LDA) [13] will be adopted, a generative statistical model widely used in different tasks in the NLP area. The LDA belongs

to the category of unsupervised machine learning algorithms, meaning that there is no human interference during the execution of the algorithm. This lack of human interaction might lead to unpleasant results, which in our case, create overlapping or -even worse- vacant claims.

A possible solution to this deficiency is the incorporation of lexical priors into the LDA [67], which are set into the model and guide it in a certain direction, and eventually, avoid the pitfalls of the unsupervised learning algorithms. With the deployment of a guided topic modelling algorithm, a previously labour-intensive and manually-consuming task has been significantly enhanced and simplified. The proposed architecture for a semi-supervised model capable of tackling the task of claim discovery is a crucial step towards the implementation of a complete AM framework, able to discover, analyze and evaluate arguments derived from noisy sources, with minimum human interference.

8.2.2 Similarity methods

The hybrid model proposed in the thesis creates a database of claims enhancing a previously manual process with unsupervised ML algorithms identifying claims in previously unseen text. However, the proposed model does not introduce an algorithm for identifying the conceptual proximity between the collected claims and the previously unseen chunks of text. For example, the claims *Trump has gone to war against California* and *LA hates Republicans* are correlated, even though they do not share any common words. In this thesis, two different approaches have been used for implementing binary representation of text: bag-of-words and TF-IDF producing sparse matrices representing text in numerical form. The calculation of the similarity of the previously "unseen" data and the stored claims -now represented as matrices- is achieved through the calculation of their cosine similarity. If the absolute value that is returned is above the defined threshold, then the chunk of text is characterized as claimed.

Alternative encoding methods could use pre-trained word embeddings such as Word2Vec, GloVe, and fastText. These methods are widely used, but never extensively tested and compared on the specific task of argumentation detection. Additionally, knowledge-based measures that quantify the semantic relatedness of words using a semantic network, such as WordNet, could provide an insight into exploiting multi-relational data that include embedding entities and relationships. Finally, the Jensen-Shannon distance could be used in conjunction with the LDA to estimate which statements are statistically "closer" (and therefore more similar) based on the argument distribution in the trained dataset.

8.2.3 Federated learning

The need to transfer knowledge between tasks and datasets is a major issue not only for the task of argumentation detection but also for the wider field of NLP. The lack of available datasets and their great heterogeneity hinders the generalization capabilities of the existing techniques and language models. Additionally, the dominant centralized approach for training the existing models on new datasets creates a weakness in expanding the existing knowledge to new domains because it requires the collection of all the new available datasets and training of a model from the beginning; a process that requires significant resources. The concept of federated learning has recently emerged, providing a partial solution to the aforementioned problems, enabling decentralised processing on heterogeneous datasets whose sizes may span

several orders of magnitude. The implementation of a federated BERT model [170] enables clients with limited computing capability to participate in pre-training a large model without forcing them to share their datasets while maintaining excellent performance. However, another research indicated that the sampling ratio is a crucial factor in the performance of the system [193], potentially creating a series of challenges in its application in real-life settings.

Federated learning has not yet been tested in the context of AM, at least to my knowledge, creating a unique opportunity to test the capabilities of a new method in a previously unexplored field. The implementation of a federated model that compiles only the findings of the local clients without requiring the raw data could diminish the hesitations on sharing datasets while it could provide a master model that could improve on every iteration, integrating the suggestions from local clients.

Bibliography

[1] Addawood, A. A. and Bashir, M. N. [2016], What is Your Evidence? A Study of Controversial Topics on Social Media, *in* 'Proceedings of the 3rd Workshop on Argument Mining', Berlin, Germany, pp. 1–11.

[2] Addawood, A., Schneider, J. and Bashir, M. [2017], Stance Classification of Twitter Debates: The Encryption Debate as A Use Case, *in* 'Proceedings of the 8th International Conference on Social Media & Society', ACM Press, New York, New York, USA, pp. 1–10.

[3] Afrin, T., Wang, E., Litman, D., Matsumura, L. C. and Correnti, R. [2020], Annotation and Classification of Evidence and Reasoning Revisions in Argumentative Writing, *in* 'Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications', Seattle, WA,, pp. 75–84.
URL: *https://www.aclweb.org/anthology/2020.bea-1.7*

[4] Apuke, O. D. and Omar, B. [2021], 'Fake news and covid-19: modelling the predictors of fake news sharing among social media users', *Telematics and Informatics* **56**, 101475.
URL: *https://www.sciencedirect.com/science/article/pii/S0736585320301349*

[5] Arsene, O., Dumitrache, I. and Mihu, I. [2015], 'Expert system for medicine diagnosis using software agents', *Expert Systems with Applications* **42**(4), 1825–1834.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417414006502*

[6] Bal, B. K. and Dizier, P. S. [2010], Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials, *in* 'Proceedings of the Seventh conference on International Language Resources and Evaluation', Valletta, Malta, pp. 1152–1158.

[7] Bar, R., Lilach, H., Friedman, E. R., Kantor, Y., Lahav, D. and Slonim, N. [2020], From Arguments to Key Points: Towards Automatic Argument Summarization, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, pp. 4029–4039.
URL: *https://arxiv.org/abs/2005.01619*

[8] Beardsley, M. C. [1950], 'Practical Logic', *The Philosophical Quarterly* .

[9] Belbachir, F. and Boughanem, M. [2018], 'Using language models to improve opinion detection', *Information Processing & Management* **54**(6), 958–968.

[10] Bex, F., Snaith, M., Lawrence, J. and Reed, C. [2014], 'ArguBlogging: An application for the Argument Web', *Web Semantics: Science, Services and Agents on the World Wide Web* **25**, 9–15.

[11] Bibal, A., Lognoul, M., Streel, A. and Frénay, B. [2021], 'Legal requirements on explainability in machine learning', *Artificial Intelligence and Law* **29**.

[12] Bird, S., Edwardm Loper and Ewan, K. [2009], *Natural Language Processing with Python*, O'Reilly Media Inc.

[13] Blei, D. M., Ng, A. Y. and Jordan, M. I. [2003], 'Latent dirichlet allocation', *J. Mach. Learn. Res.* **3**(null), 993–1022.

[14] Boltužić, F. and Šnajder, J. [2014], Back up your Stance: Recognizing Arguments in Online Discussions, *in* 'Proceedings of the First Workshop on Argumentation Mining', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 49–58.

[15] Boltužic, F. and Snajder, J. [2016], Fill the Gap! Analyzing Implicit Premises between Claims from Online Debates, *in* 'Proceedings of the 3rd Workshop on Argument Mining', Berlin, Germany, pp. 124–133.

[16] Bontcheva, K., Cunningham, H., Roberts, I., Roberts, A., Tablan, V., Aswani, N. and Gorrell, G. [2013], 'GATE Teamware: a web-based, collaborative text annotation framework', *Language Resources and Evaluation* **47**(4), 1007–1029.

[17] Bosc, T., Cabrio, E. and Villata, S. [2016*a*], DART: a Dataset of Arguments and their Relations on Twitter - Semantic Scholar, *in* 'LREC', Portorož, Slovenia, pp. 1258–1263.

[18] Bosc, T., Cabrio, E. and Villata, S. [2016*b*], Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media., *in* 'Proceedings of the 6th International Conference on Computational Models of Argument', Potsdam, Germany, pp. 21–32.

[19] Brown, D. K., Ng, Y. M. M., Riedl, M. J. and Lacasa-Mas, I. [2018], 'Reddit's veil of anonymity: Predictors of engagement and participation in media environments with hostile reputations', *Social Media + Society* **4**(4), 2056305118810216.
URL: *https://doi.org/10.1177/2056305118810216*

[20] Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. [1995], 'A Limited Memory Algorithm for Bound Constrained Optimization', *SIAM Journal on Scientific Computing* **16**(5), 1190–1208.

[21] Cabrio, E. and Villata, S. [2018], Five Years of Argument Mining: a Data-driven Analysis, *in* 'Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence', International Joint Conferences on Artificial Intelligence Organization, California, pp. 5427–5433.

[22] Carstens, L. and Toni, F. [2017], 'Using Argumentation to Improve Classification in Natural Language Problems', *ACM Transactions on Internet Technology* **17**(3), 1–23.

[23] Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K. and Hwang, A. [2019], AMPERSAND: Argument Mining for PERSuAsive oNline Discussions, *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)', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 2933–2943.
URL: *https://www.aclweb.org/anthology/D19-1291*

[24] Chandra Pandey, A., Singh Rajpoot, D. and Saraswat, M. [2017], 'Twitter sentiment analysis using hybrid cuckoo search method', *Information Processing & Management* **53**(4), 764–779.

[25] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. [2014], Learning phrase representations using RNN encoder-decoder for statistical machine translation, *in* 'EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference', Association for Computational Linguistics (ACL), pp. 1724–1734.
URL: *https://arxiv.org/abs/1406.1078v3*

[26] Clifton, A., Pappu, A., Reddy, S., Yu, Y., Karlgren, J., Carterette, B. and Jones, R. [2020], 'The spotify podcast dataset', *arXiv* .
URL: *https://arxiv.org/abs/2004.04270*

[27] Cocarascu, O. and Toni, F. [2018], 'Combining deep learning and argumentative reasoning for the analysis of social media textual content using small datasets', *Computational Linguistics* pp. 1–37.

[28] Cohen, J. [1960], 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* **20**(1), 37–46.
URL: *https://doi.org/10.1177/001316446002000104*

[29] Cortis, K., Freitas, A., Daudert, T., Hürlimann, M., Zarrouk, M., Handschuh, S. and Davis, B. [2017], SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News, *in* 'Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)', Vancouver, Canad, pp. 519–535.
URL: *http://www.aclweb.org/anthology/S17-2089*

[30] Cunningham, H., Tablan, V., Roberts, A. and Bontcheva, K. [2013], 'Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics', *PLoS Computational Biology* **9**(2), e1002854.

[31] Dass, S. D. S., Meskaran, F. and Saeedi, M. [2020], 'Expert system for early diagnosis of covid-19', *International Journal of Current Research and Review* **12**, 162–165.

[32] Deturck, K., Nouvel, D. and Segond, F. [2018], ERTIM@MC2: Diversified Argumentative Tweets Retrieval, *in* 'CLEF MC2 2018 Lab Overview', Avignon, France, pp. 302–308.

[33] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. [2018], BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *in* 'NAACL HLT 2019 -

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference', Vol. 1, Association for Computational Linguistics (ACL), pp. 4171–4186.
URL: *http://arxiv.org/abs/1810.04805*

[34] Dufour, R., Mickael, R., Delorme, A. and Malinas, D. [2018], LIA@CLEF 2018: Mining Events Opinion Argumentation from Raw Unlabeled Twitter Data using Convolutional Neural Network., *in* 'Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum', Avignon, France.

[35] Dung, P. M. [1995], 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial Intelligence* **77**(2), 321–357.

[36] Dusmanu, M., Cabrio, E. and Villata, S. [2017], Argument Mining on Twitter: Arguments, Facts and Sources, *in* 'Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing', Copenhagen, Denmark, pp. 2317–2322.

[37] Ebrahimi, J., Dou, D. and Lowd, D. [2016], Weakly Supervised Tweet Stance Classification by Relational Bootstrapping, *in* 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', Austin, Texas, p. 1012–1017.

[38] Eckart De Castilho, R., Ujdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A. and Biemann, C. [2016], A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures, *in* 'Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)', Osaka, Japan, pp. 76–84.

[39] Eidelman, V. and Grom, B. [2019], 'Argument identification in public comments from erulemaking', *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* .

[40] Eryiit, U., Samet, F., Etin, C., Yanık, M. and Temel, T. [2013], TURKSENT: A Sentiment Annotation Tool for Social Media, *in* 'Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse', Sofia, Bulgaria, pp. 131–134.

[41] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J. [2008], 'Liblinear: A library for large linear classification', *J. Mach. Learn. Res.* **9**, 1871–1874.

[42] Fischer, S. [2017], 'Lost in regulation: The EU and Nord Stream 2', *CSS Policy Perspectives* **5**(5).
URL: *https://doi.org/10.3929/ethz-b-000210438*

[43] Freeman, J. B. [1991], *Dialectics and the macrostructure of arguments: a theory of argument structure*, Foris Publications.

[44] Freeman, J. B. [2011], *Argument structure : representation and theory*, Springer.

[45] Fromm, M., Faerman, E. and Seidl, T. [2019], 'TACAM: Topic And Context Aware Argument Mining', *Proceedings - 2019 IEEE/WIC/ACM International Conference on*

Web Intelligence, WI 2019 pp. 99–106.
URL: *http://arxiv.org/abs/1906.00923 http://dx.doi.org/10.1145/3350546.3352506*

[46] Galassi, A., Lippi, M. and Torroni, P. [2018], Argumentative Link Prediction using Residual Networks and Multi-Objective Learning, *in* 'Proceedings of the 5th Workshop on Argument Mining', Brussels, Belgium, pp. 1–10.

[47] Gangal, V. and Hovy, E. [2020], 'BERTering RAMS: What and How Much does BERT Already Know About Event Arguments? – A Study on the RAMS Dataset'.
URL: *http://arxiv.org/abs/2010.04098*

[48] Gelder, T. V. [2000], Learning to reason: a Reason!-Able approach, *in* 'Cognitive Science in Australia, 2000: Proceedings of the Fifth Australasian Cognitive Science Society Conference', Adelaide.

[49] Giachanou, A. and Crestani, F. [2016], 'Like It or Not', *ACM Computing Surveys* **49**(2), 1–41.

[50] Gil, R. and Lopes, C. H. [2022], Context matters! identifying argumentative relations in essays, *in* 'The 37th ACM/SIGAPP Symposium On Applied Computing'.

[51] Goudas, T., Louizos, C., Petasis, G. and Karkaletsis, V. [2014], Argument Extraction from News, Blogs, and Social Media, *in* 'Artificial Intelligence: Methods and Applications.SETN 2014.', Springer, Cham, pp. 287–299.

[52] Grasso, F., Cawsey, A. and Jones, R. [2000], 'Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition', *International Journal of Human-Computer Studies* **53**(6), 1077–1115.

[53] Grčar, M., Cherepnalkoski, D., Mozetič, I. and Kralj Novak, P. [2017], 'Stance and influence of Twitter users regarding the Brexit referendum', *Computational Social Networks* **4**(1), 6.

[54] Green, N. L. [2017], Manual Identification of Arguments with Implicit Conclusions Using Semantic Rules for Argument Mining, *in* 'Proceedings of the 4th Workshop on Argument Mining', Copenhagen, Denmark, pp. 73–78.

[55] Green, N. L. [2018], 'Towards mining scientific discourse using argumentation schemes', *Argument & Computation* **9**(2), 121–135.
URL: *http://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/AAC-180038*

[56] Guo, J., Chen, T. and Wu, W. [2021], 'A multi-feature diffusion model: Rumor blocking in social networks', *IEEE/ACM Transactions on Networking* **29**(1), 386–397.

[57] Guo, W., Li, H., Ji, H. and Diab, M. [2013], Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media, *in* 'Proceedings of the 51st Annual Meeting of the Association for Computational Linguists', Sofia, Bulgaria, p. 239–249.

[58] Habernal, I. and Gurevych, I. [2017*a*], 'Argumentation Mining in User-Generated Web Discourse', *Computational Linguistics* **43**(1), 125–179.
URL: *https://doi.org/10.1162/COLI_a_00276*

[59] Habernal, I. and Gurevych, I. [2017*b*], 'Argumentation Mining in User-Generated Web Discourse', *Computational Linguistics* **43**(1), 125–179.

[60] Habernal, I., Wachsmuth, H., Gurevych, I. and Stein, B. [2018], The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants, *in* '16th North American Chapter of the Association for Computational Linguistics: Human Language Technologies', New Orleans, Louisiana, USA, pp. 1930–1940.

[61] Hashemi, M. [2017], 'The infrastructure behind twitter: Scale'.
URL: *https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale*

[62] Hosni, A. I. E., Li, K. and Ahmed, S. [2018], Hisbmodel: A rumor diffusion model based on human individual and social behaviors in online social networks, *in* L. Cheng, A. C. S. Leung and S. Ozawa, eds, 'Neural Information Processing', Springer International Publishing, Cham, pp. 14–27.

[63] Howard, J. and Ruder, S. [2018], 'Universal Language Model Fine-tuning for Text Classification', *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* **1**, 328–339.
URL: *http://arxiv.org/abs/1801.06146*

[64] Hu, D. [2020], An introductory survey on attention mechanisms in nlp problems, *in* Y. Bi, R. Bhatia and S. Kapoor, eds, 'Intelligent Systems and Applications', Springer International Publishing, Cham, pp. 432–448.

[65] Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S. and Fernández, S. [2016], A twitter sentiment gold standard for the brexit referendum, *in* 'Proceedings of the 12th International Conference on Semantic Systems', SEMANTiCS 2016, Association for Computing Machinery, New York, NY, USA, p. 193–196.
URL: *https://doi.org/10.1145/2993318.2993350*

[66] Introne, J., Yildirim, I. G., Iandoli, L., DeCook, J. and Elzeini, S. [2018], 'How people weave online information into pseudoknowledge', *Social Media + Society* **4**(3), 2056305118785639.
URL: *https://doi.org/10.1177/2056305118785639*

[67] Jagarlamudi, J., Daumé III, H. and Udupa, R. [2012], Incorporating lexical priors into topic models, *in* 'Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Avignon, France, pp. 204–213.
URL: *https://aclanthology.org/E12-1021*

[68] Jaidka, K., Chhaya, N. and Ungar, L. [2018], Diachronic degradation of language models: Insights from social media, *in* 'Proceedings of the 56th Annual Meeting of the Association

for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 195–200.
URL: *https://aclanthology.org/P18-2032*

[69] Janier Mathilde, Lawrence John and Reed Chris [2014], OVA+: An argument analysis interface, *in* 'Proceedings of the 5th International Conference on Computational Models of Argument (COMMA'14).', p. 463–464.

[70] Jennings, N., Faratin, P., Lomuscio, A., Parsons, S., Wooldridge, M. and Sierra, C. [2001], 'Automated Negotiation: Prospects, Methods and Challenges', *Group Decision and Negotiation* **10**(2), 199–215.

[71] Jing, E., Schneck, K. E., Egan, D. and Waterman, S. A. [2021], Identifying introductions in podcast episodes from automatically generated transcripts, *in* 'CM Conference on Recommender Systems (RecSys 2021)', Vol. abs/2110.07096.

[72] Johnson, K. and Goldwasser, D. [2016], Identifying Stance by Analyzing Political Discourse on Twitter, *in* 'Proceedings of the First Workshop on NLP and Computational Social Science', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 66–75.

[73] Jones, R., Zamani, H., Schedl, M., Chen, C.-W., Reddy, S., Clifton, A., Karlgren, J., Hashemi, H., Pappu, A., Nazari, Z., Yang, L., Semerci, O., Bouchard, H. and Carterette, B. [2021], *Current Challenges and Future Directions in Podcast Information Access*, Association for Computing Machinery, New York, NY, USA, p. 1554–1565.
URL: *https://doi.org/10.1145/3404835.3462805*

[74] Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. M. and Wu, Y. [2016], 'Exploring the limits of language modeling', *ArXiv* **abs/1602.02410**.

[75] Karimi, M., Jannach, D. and Jugovac, M. [2018], 'News recommender systems – Survey and roads ahead', *Information Processing & Management* **54**(6), 1203–1227.

[76] Khan, F. H., Bashir, S. and Qamar, U. [2014], 'TOM: Twitter opinion mining framework using hybrid classification scheme', *Decision Support Systems* **57**, 245–257.

[77] Kirschner, C., Eckle-Kohler, J. and Gurevych, I. [2015], Linking the Thoughts: Analysis of Argumentation Structures in Scientific Publications, *in* 'Proceedings of the 2nd Workshop on Argumentation Mining', Denver, Colorado, pp. 1–11.

[78] Konstantinovskiy, L., Price, O., Babakar, M. and Zubiaga, A. [2018], Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection, *in* 'EMNLP 2018: Conference on Empirical Methods in Natural Language Processing', Brussels, Belgium.

[79] Krause, P., Ambler, S., Elvang-Goransson, M. and Fox, J. [1995], 'A Logic of Argumentation for Reasoning under Uncertainty', *Computational Intelligence* **11**(1), 113–131.

[80] Krippendorff, K. [1980], *Content analysis: an introduction to its methodology*, SAGE publications.

[81] Kuribayashi, T., Reisert, P., Inoue, N. and Inui, K. [2018], Towards Exploiting Argumentative Context for Argumentative Relation Identification, *in* 'Proceedings of the 24th Annual Conference of the Society of Language Processing (March 2018)', pp. 284–287.

[82] Kurtanović, Z. and Maalej, W. [2018], 'On user rationale in software engineering', *Requirements Engineering* pp. 1–23.

[83] Lai, M., Patti, V., Ruffo, G. and Rosso, P. [2018], Stance Evolution and Twitter Interactions in an Italian Political Debate, *in* 'NLDB 2018: Natural Language Processing and Information Systems', Springer, Cham, Paris, France, pp. 15–27.

[84] Lai, M., Tambuscio, M., Patti, V., Ruffo, G. and Rosso, P. [2017], Extracting Graph Topological Information and Users' Opinion, *in* 'Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017. Lecture Notes in Computer Science', Springer, Cham, pp. 112–118.
URL: *http://link.springer.com/10.1007/978-3-319-65813-1_10*

[85] Landis, J. R. and Koch, G. G. [1977], 'The measurement of observer agreement for categorical data.', *Biometrics* **33**(1), 159–74.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/843571*

[86] Lauscher, A., Glavaš, G. and Eckert, K. [2018], ArguminSci: A Tool for Analyzing Argumentation and Rhetorical Aspects in Scientific Writing, *in* 'Proceedings of the 5th Workshop on Argument Mining', Association for Computational Linguistics, Brussels, Belgium, pp. 22–28.
URL: *https://aclweb.org/anthology/papers/W/W18/W18-5203/*

[87] Lauscher, A., Glavaš, G. and Ponzetto, S. P. [2018], An Argument-Annotated Corpus of Scientific Publications, *in* 'Proceedings of the 5th Workshop on Argument Mining', Association for Computational Linguistics, Brussels, Belgium, pp. 40–46.
URL: *https://aclweb.org/anthology/papers/W/W18/W18-5206/*

[88] Lawrence, J., Park, J., Budzynska, K., Cardie, C., Konat, B. and Reed, C. [2017], 'Using Argumentative Structure to Interpret Debates in Online Deliberative Democracy and eRulemaking', *ACM Transactions on Internet Technology* **17**(3), 1–22.

[89] Lawrence, J., Snaith, M., Konat, B., Budzynska, K. and Reed, C. [2017], 'Debating Technology for Dialogical Argument', *ACM Transactions on Internet Technology* **17**(3), 1–23.

[90] Lee, S., Ha, T., Lee, D. and Kim, J. H. [2018], 'Understanding the majority opinion formation process in online environments: An exploratory approach to Facebook', *Information Processing & Management* **54**(6), 1115–1128.

[91] Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E. and Slonim, N. [2014], Context Dependent Claim Detection, *in* 'COLING - International Committee on Computational Linguistics', Dublin, Ireland, pp. 1489–1500.

[92] Liebeck, M., Esau, K. and Conrad, S. [2016], What to Do with an Airport? Mining Arguments in the German Online Participation Project Tempelhofer Feld, *in* 'Proceedings of the Third Workshop on Argument Mining (ArgMining2016)', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 144–153.

[93] Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G. and Torroni, P. [2018], 'CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service', *arXiv preprint* .

[94] Lippi, M. and Torroni, P. [2015], Context-independent claim detection for argument mining, *in* 'Proceedings of the 24th International Conference on Artificial Intelligence', AAAI Press = The Association for the Advancement of Artificial Intelligence Press, Buenos Aires, Argentina, pp. 185–191.

[95] Lippi, M. and Torroni, P. [2016*a*], Argument Mining from Speech: Detecting Claims in Political Debates, *in* 'AAAI'.

[96] Lippi, M. and Torroni, P. [2016*b*], 'Argumentation mining: State of the art and emerging trends', *ACM Transactions on Internet Technology* **16**(2), 1–25.

[97] Lippi, M. and Torroni, P. [2016*c*], 'MARGOT: A web server for argumentation mining', *Expert Systems with Applications* **65**, 292–303.

[98] Liu, B. [2012], *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers.

[99] Lytos, A., Lagkas, T., Sarigiannidis, P. and Bontcheva, K. [2018], Argumentation Mining: Exploiting Multiple Sources and Background Knowledge, *in* '12th South East European Doctoral Student Conference DSC2018'.
URL: *https://www.researchgate.net/publication/327728501*

[100] Lytos, A., Lagkas, T., Sarigiannidis, P. and Bontcheva, K. [2019], 'The evolution of argumentation mining: From models to social media and emerging tools', *Information Processing & Management* **56**(6), 102055.
URL: *http://www.sciencedirect.com/science/article/pii/S030645731930024X*

[101] Ma, W., Chao, W., Luo, Z. and Jiang, X. [2018], Claim Retrieval in Twitter, *in* 'Web Information Systems Engineering – WISE 2018', Dubai, United Arab Emirates, pp. 297–307.

[102] Mann, W. C. [1984], Discourse structures for text generation, *in* 'Proceedings of the 10th international conference on Computational linguistics -', Association for Computational Linguistics, Morristown, NJ, USA, pp. 367–375.

[103] Mann, W. C. and Thompson, S. A. [1988], 'Rhetorical Structure Theory: Toward a functional theory of text organization', *Text - Interdisciplinary Journal for the Study of Discourse* **8**(3), 243–281.

[104] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. [2014], The Stanford CoreNLP Natural Language Processing Toolkit, *in* 'Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations', Association for Computational Linguistics, Baltimore, Maryland, pp. 55–60.

[105] Marriott, K., Sbarski, P., van Gelder, T., Prager, D. and Bulka, A. [2011], 'Hi-Trees and Their Layout', *IEEE Transactions on Visualization and Computer Graphics* **17**(3), 290–304.

[106] Mayer, T., Cabrio, E. and Villata, S. [2020], Transformer-based Argument Mining for Healthcare Applications, *in* '24th European Conference on Artificial Intelligence (ECAI2020)', Santiago de Compostela, Spain.
URL: *https://hal.archives-ouvertes.fr/hal-02879293/document*

[107] Maynard, D., Roberts, I., Greenwood, M. A., Rout, D. and Bontcheva, K. [2017], 'A framework for real-time semantic social media analysis', *Web Semantics: Science, Services and Agents on the World Wide Web* **44**, 75–88.

[108] Mikolov, T., Chen, K., Corrado, G. and Dean, J. [2013], Efficient Estimation of Word Representations in Vector Space, *in* 'Proceedings of Workshop at ICLR'.
URL: *https://arxiv.org/abs/1301.3781*

[109] Mochales, R. and Moens, M.-F. [2011], 'Argumentation mining', *Artificial Intelligence and Law* **19**(1), 1–22.

[110] Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X. and Cherry, C. [2016], SemEval-2016 Task 6: Detecting Stance in Tweets, *in* 'International Workshop on Semantic Evaluation (SemEval-2016)', San Diego, California, pp. 31–41.

[111] Mohammad, S. M., Sobhani, P. and Kiritchenko, S. [2017], 'Stance and Sentiment in Tweets', *ACM Transactions on Internet Technology* **17**(3), 1–23.

[112] Morio, G. and Fujita, K. [2018], Annotating Online Civic Discussion Threads for Argument Mining, *in* 'Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2018 (WI'18)', IEEE, Santiago, Chile, pp. 801–807.

[113] Naderi, N. and Hirst, G. [2015], Argumentation Mining in Parliamentary Discourse, *in* 'Workshop on Computational Models of Natural Argument, International Workshop on Empathic Computing', Springer, Cham, Bertinoro, Italy, pp. 16–25.

[114] Naderi, N. and Hirst, G. [2016], Argumentation mining in parliamentary discourse, *in* M. Baldoni, C. Baroglio, F. Bex, F. Grasso, N. Green, M.-R. Namazi-Rad, M. Numao and M. T. Suarez, eds, 'Principles and Practice of Multi-Agent Systems', Springer International Publishing, Cham, pp. 16–25.

[115] Nguyen, H. T. and Le Nguyen, M. [2018], 'Multilingual opinion mining on YouTube – A convolutional N-gram BiLSTM word embedding', *Information Processing & Management* **54**(3), 451–462.

[116] O'Donnell, M. [2000], RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory, *in* 'Proceedings of the International Natural Language Generation Conference (INLG'2000)', Mitzpe Ramon, Israel, pp. 253 – 256.

[117] Ollinger, S., Dumani, L., Sahitaj, P., Bergmann, R. and Schenkel, R. [2020], 'Same Side Stance Classification Task: Facilitating Argument Stance Classification by Fine-tuning a BERT Model', *arXiv* .
URL: *http://arxiv.org/abs/2004.11163*

[118] Ott, B. L. [2017], 'The age of twitter: Donald j. trump and the politics of debasement', *Critical Studies in Media Communication* **34**(1), 59–68.

[119] Palau, R. M. and Moens, M.-F. [2009], Argumentation mining: The detection, classification and structure of arguments in text, *in* 'Proceedings of the 12th International Conference on Artificial Intelligence and Law', ICAIL '09, Association for Computing Machinery, New York, NY, USA, p. 98–107.
URL: *https://doi.org/10.1145/1568234.1568246*

[120] Park, J. and Cardie, C. [2014], Identifying Appropriate Support for Propositions in Online User Comments, *in* 'Proceedings of the First Workshop on Argumentation Mining', Baltimore, Maryland USA, pp. 29–38.

[121] Park, J., Katiyar, A. and Yang, B. [2015], Conditional Random Fields for Identifying Appropriate Types of Support for Propositions in Online User Comments, *in* 'Proceedings of the 2nd Workshop on Argumentation Mining', Denver, Colorado, pp. 39–44.

[122] Parsons, S. D. and Jennings, N. R. [1996], Neogotiation Through Argumentation - A Preliminary Report, *in* '2nd Int. Conf. on Multi-Agent Systems', Japan, pp. 267–274.

[123] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. [2011], 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research* **12**(Oct), 2825–2830.
URL: *http://www.jmlr.org/papers/v12/pedregosa11a.html*

[124] Peldszus, A. and Stede, M. [2013], 'From Argument Diagrams to Argumentation Mining in Texts', *International Journal of Cognitive Informatics and Natural Intelligence* **7**(1), 1–31.

[125] Peldszus, A. and Stede, M. [2015], Joint prediction in MST-style discourse parsing for argumentation mining, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 938–948.

[126] Peldszus, A. and Stede, M. [2016], Rhetorical structure and argumentation structure in monologue text, *in* 'Proceedings of the 3rd Workshop on Argument Mining', Berlin, Germany, pp. 103–112.

[127] Pennebaker, J. W. [1997], 'Writing About Emotional Experiences as a Therapeutic Process', *Psychological Science* **8**(3), 162–166.

[128] Pennebaker, J. W. and Francis, M. E. [1996], 'Cognitive, Emotional, and Language Processes in Disclosure', *Cognition and Emotion* **10**(6), 601–626.

[129] Pennington, J., Socher, R. and Manning, C. [2014], Glove: Global Vectors for Word Representation, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1532–1543.
URL: *http://aclweb.org/anthology/D14-1162*

[130] Pereira, M. H. R., Pádua, F. L. C., Pereira, A. M., Benevenuto, F. and Dalip, D. H. [2016], Fusing audio, textual, and visual features for sentiment analysis of news videos, *in* 'ICWSM'.

[131] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. [2018], Deep contextualized word representations, *in* 'NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference', Vol. 1, Association for Computational Linguistics (ACL), pp. 2227–2237.
URL: *http://allennlp.org/elmo*

[132] Pollock, J. L. [1987], 'Defeasible reasoning', *Cognitive Science* **11**(4), 481–518.

[133] Pollock, J. L. [2001], 'Defeasible reasoning with variable degrees of justification', *Artificial Intelligence* **133**(1-2), 233–282.

[134] Poria, S., Cambria, E., Howard, N., Huang, G.-B. and Hussain, A. [2016], 'Fusing audio, visual and textual clues for sentiment analysis from multimodal content', *Neurocomputing* **174**, 50–59.
URL: *https://www.sciencedirect.com/science/article/pii/S0925231215011297*

[135] Radford, A. and Narasimhan, K. [2018], Improving Language Understanding by Generative Pre-Training, *in* 'Pre-print'.

[136] Rajendran, P. [2016], Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews, *in* 'Proceedings of the 3rd Workshop on Argument Mining', Berlin, Germany, pp. 31–39.

[137] Rajendran, P., Bollegala, D. and Parsons, S. [2016], Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews, *in* 'Proceedings of the Third Workshop on Argument Mining (ArgMining2016)', Association for Computational Linguistics, Berlin, Germany, pp. 31–39.
URL: *https://aclanthology.org/W16-2804*

[138] Rajendran, P., Bollegala, D. and Parsons, S. [2018], Is Something Better than Nothing? Automatically Predicting Stance-based Arguments using Deep Learning and Small Labelled Dataset, *in* '16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', New Orleans, Louisiana, pp. 28–34.

[139] Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A. and Snaith, M. [2017], 'The Argument Web: an Online Ecosystem of Tools, Systems and Services for Argumentation', *Philosophy & Technology* **30**(2), 137–160.

[140] Reed, C. and Rowe, G. [2004], 'Araucaria: Software For Argument Analysis, Diagramming And Representation', *International Journal on Artificial Intelligence Tools* **13**(04), 961–979.

[141] Reed, C., Walton, D. and Macagno, F. [2007], 'Argument diagramming in logic, law and artificial intelligence', *The Knowledge Engineering Review* **22**(01), 87.

[142] Reed, C. and Wells, S. [2007], 'Dialogical Argument as an Interface to Complex Debates', *IEEE Intelligent Systems* **22**(6), 60–65.

[143] Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C. and Gurevych, I. [2019], Classification and Clustering of Arguments with Contextualized Word Embeddings, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 567–578.
URL: *https://www.aclweb.org/anthology/P19-1054*

[144] Rinott, R., Dankin, L., Alzate Perez, C., Khapra, M. M., Aharoni, E. and Slonim, N. [2015], Show Me Your Evidence - an Automatic Method for Context Dependent Evidence Detection, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 440–450.

[145] Roitman, H., Hummel, S., Rabinovich, E., Sznajder, B., Slonim, N. and Aharoni, E. [2016], On the Retrieval of Wikipedia Articles Containing Claims on Controversial Topics, *in* 'Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion', ACM Press, New York, New York, USA, pp. 991–996.

[146] Rowland, A., Craig-Hare, J., Ault, M., Ellis, J. and Bulgren, J. [2017], 'Social media: How the next generation can practice argumentation', *Educational Media International* **54**(2), 99–111.
URL: *https://doi.org/10.1080/09523987.2017.1362818*

[147] Rufai, S. R. and Bunce, C. [2020], 'World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis', *Journal of Public Health* **42**(3), 510–516.
URL: *https://doi.org/10.1093/pubmed/fdaa049*

[148] Saint-Dizier, P. [2012], 'Processing natural language arguments with the TextCoop platform', *Argument & Computation* **3**(1), 49–82.

[149] Sardianos, C., Katakis, I. M., Petasis, G. and Karkaletsis, V. [2015], Argument Extraction from News, *in* 'Proceedings of the 2nd Workshop on Argumentation Mining', Denver, Colorado, pp. 56–66.

[150] Savelka, J. and Ashley, K. D. [2016], Extracting Case Law Sentences for Argumentation about the Meaning of Statutory Terms, *in* 'Proceedings of the 3rd Workshop on Argument Mining', pp. 50–59.

[151] Sbarski, P., van Gelder, T., Marriott, K., Prager, D. and Bulka, A. [2008], Visualizing Argument Structure, *in* 'International Symposium on Visual Computing 2008: Visualizing Argument Structure', Springer, Berlin, Heidelberg, Las Vegas, NV, USA, pp. 129–138.

[152] Schaefer, R. and Stede, M. [2019], Improving Implicit Stance Classification in Tweets Using Word and Sentence Embeddings, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 11793 LNAI, Springer Verlag, pp. 299–307.

[153] Schulz, C., Eger, S., Daxenberger, J., Kahse, T. and Gurevych, I. [2018], Multi-Task Learning for Argumentation Mining in Low-Resource Settings, *in* 'NAACL HLT 2018', pp. 35–41.

[154] Sendi, S. and Latiri, C. [2018], Opinion Argumentation based on Combined Information Retrieval and Topic Modeling., *in* 'Working Notes of CLEF 2018 - Conference and Labs of the Forum', Avignon, France.

[155] Skeppstedt, M., Sahlgren, M., Paradis, C. and Kerren, A. [2016], Unshared task: (Dis)agreement in online debates, *in* 'Proceedings of the 3rd Workshop on Argument Mining', Berlin, Germany, pp. 154–159.

[156] Snaith, M., Lawrence, J. and Reed, C. [2010], Mixed Initiative Argument in Public Deliberation, *in* 'International Conference on Online Deliberation', Leeds, UK, pp. 2–13.

[157] Snaith Mark and Reed Chris [2012], TOAST: Online ASPIC+ implementation, *in* 'Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)', Vienna, Austria.

[158] Sonntag, J. and Stede, M. [2014], GraPAT: a Tool for Graph Annotations, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)', Reykjavik, Iceland, pp. 4141–4151.

[159] Stab, C., Daxenberger, J., Stahlhut, C., Miller, T., Schiller, B., Tauchmann, C., Eger, S. and Gurevych, I. [2018], ArgumenText: Searching for Arguments in Heterogeneous Sources, *in* 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 21–25.

[160] Stab, C. and Gurevych, I. [2014], Annotating Argument Components and Relations in Persuasive Essays, *in* 'Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers ,', Dublin, Ireland, pp. 1501–1510.

[161] Stab, C., Miller, T. and Gurevych, I. [2018], 'Cross-topic Argument Mining from Heterogeneous Sources Using Attention-based Neural Networks', *CoRR* .
URL: *https://dblp.uni-trier.de/rec/bibtex/journals/corr/abs-1802-05758*

[162] Stella, P. [2017], 'A deeper dive into reddit's design patterns and information architecture'.
URL: *https://uxplanet.org/a-deeper-dive-into-reddits-design-patterns-and-information-architecture-634eae96a6*

[163] Stenetorp, P., Pyysalo, S., Topi, G., Ohta, T., Ananiadou, S. and Tsujii, J. i. [2012], BRAT: a Web-based Tool for NLP-Assisted Text Annotation, *in* 'Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics', Avignon, France, pp. 102–107.

[164] Stenetorp, P., Topić, G., Pyysalo, S., Ohta, T., Kim, J.-D. and Tsujii, J. [2011], BioNLP Shared Task 2011: supporting resources, *in* 'Proceedings of the BioNLP Shared Task 2011 Workshop', Association for Computational Linguistics, Portland, Oregon, pp. 112–120.

[165] Steven Loria [n.d.], 'TextBlob: Simplified Text Processing — TextBlob 0.15.1 documentation'.

[166] Stolee, G. and Caton, S. [2018], 'Twitter, trump, and the base: A shift to a new form of presidential talk?', *Signs and Society* **6**(1), 147–165.
URL: *https://doi.org/10.1086%2F694755*

[167] Subhabrata, D., Jeevesh, J., Dipankar, D. and Tanmoy, C. [2022], 'Can unsupervised knowledge transfer from social discussions help argument mining?', *arXiv* .
URL: *https://arxiv.org/abs/2203.12881v1*

[168] Tan, S. [2017], Spot the lie: Detecting untruthful online opinion on twitter., PhD thesis, Imperial College London.

[169] *The Architecture Twitter Uses to Deal with 150M Active Users, 300K QPS, a 22 MB/S Firehose, and Send Tweets in Under 5 Seconds* [2013].
URL: *http://highscalability.com/blog/2013/7/8/the-architecture-twitter-uses-to-deal-with-150m-active-users.html*

[170] Tian, Y., Wan, Y., Lyu, L., Yao, D., Jin, H. and Sun, L. [2022], '¡span class="smallcaps smallercapital"¿fedbert¡/span¿: When federated learning meets pre-training', *ACM Trans. Intell. Syst. Technol.* .
URL: *https://doi.org/10.1145/3510033*

[171] Toledo-Ronen, O., Orbach, M., Bilu, Y., Spector, A. and Slonim, N. [2020], 'Multilingual Argument Mining: Datasets and Analysis', pp. 303–317.
URL: *http://arxiv.org/abs/2010.06432*

[172] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G. and van Moorsel, A. [2020], The relationship between trust in ai and trustworthy machine learning technologies, *in* 'FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency', FAT* '20, Association for Computing Machinery, New York, NY, USA, p. 272–283.
URL: *https://doi.org/10.1145/3351095.3372834*

[173] Toulmin, S. E. [1958], *The Uses of Argument*, Cambridge University Press.

[174] Toulmin, S. E. [2003], *The Uses of Argument*, Cambridge University Press, Cambridge.

[175] Tubishat, M., Idris, N. and Abushariah, M. A. [2018], 'Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges', *Information Processing & Management* **54**(4), 545–563.

[176] van Gelder, T. [2007], 'The rationale for RationaleTM', *Law, Probability and Risk* **6**(1-4), 23–42.

[177] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. [2017], 'Attention is All You Need'.
URL: *https://research.google/pubs/pub46201/*

[178] Villena-Román, J., Collada-Pérez, S., Lana-Serrano, S. and González, J. [2011], Hybrid approach combining machine learning and a rule-based expert system for text categorization, *in* 'FLAIRS Conference'.

[179] Wachsmuth, H., Kiesel, J. and Stein, B. [2015], Sentiment Flow - A General Model of Web Review Argumentation, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', Lisbon, Portugal, pp. 601–611.

[180] Wachsmuth, H., Potthast, M., Al Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J. and Stein, B. [2017], Building an Argument Search Engine for the Web, *in* 'Proceedings of the 4th Workshop on Argument Mining', Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 49–59.

[181] Wachsmuth, H. and Stein, B. [2017], 'A Universal Model for Discourse-Level Argumentation Analysis', *ACM Transactions on Internet Technology* **17**(3), 1–24.

[182] Walton, D. [2011], 'How to Refute an Argument Using Artificial Intelligence', *Studies in Logic, Grammar and Rhetoric* **23**(36), 123–154.

[183] Walton, D. and Macagno, F. [2016], 'A classification system for argumentation schemes', *Argument & Computation* **6**(3), 219–245.
URL: *http://content.iospress.com/doi/1080/19462166.2015.1123772*

[184] Wei, P., Lin, J. and Mao, W. [2018], Multi-Target Stance Detection via a Dynamic Memory-Augmented Network, *in* 'The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18', ACM Press, New York, New York, USA, pp. 1229–1232.

[185] Wei, W., Zhang, X., Liu, X., Chen, W. and Wang, T. [2016], pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection, *in* 'Proceedings of SemEval-2016', San Diego, California, pp. 384–388.

[186] Whately, R. [1857], *Elements of logic.*, Harper & Brothers, New York, USA.

[187] Wojatzki, M. M. and Zesch, T. [2016], Stance-based Argument Mining - Modeling Implicit Argumentation Using Stance, *in* 'Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)', Bochum, Germany, pp. 313–322.

[188] Xiao, Y., Yang, Q., Sang, C. and Liu, Y. [2020], 'Rumor diffusion model based on representation learning and anti-rumor', *IEEE Transactions on Network and Service Management* **17**(3), 1910–1923.

[189] Xinyu, W., Yohan, L. and Juneyoung, P. [2022], 'Automated evaluation for student argumentative writing: A survey', *arXiv* .
URL: *https://arxiv.org/abs/2205.04083*

[190] Yimam, S. M., Gurevych, I., Eckart De Castilho, R. and Biemann, C. [2013], WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations, *in* 'Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics', Sofia, Bulgaria, pp. 1–6.

[191] Zarrella, G. and Marsh, A. [2016], MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection, *in* 'International Workshop on Semantic Evaluation (SemEval-2016)', San Diego, California, p. 458–463.

[192] Zhang, G., Nulty, P. and Lillis, D. [2022], 'Enhancing legal argument mining with domain pre-training and neural networks', *CoRR* **abs/2202.13457**.
URL: *https://arxiv.org/abs/2202.13457*

[193] Zhu, X., Wang, J., Hong, Z. and Xiao, J. [2020], Empirical studies of institutional federated learning for natural language processing, *in* 'Findings of the Association for Computational Linguistics: EMNLP 2020', Association for Computational Linguistics, Online, pp. 625–634.
URL: *https://aclanthology.org/2020.findings-emnlp.55*

Appendices

Appendix A

Emerging tools in AM

A.1 General-purpose NLP tools

The increasing interest in AM has increased the need for suitable tools, such as grammar parsers, sentiment lexicons, software for boosting the manual annotation tasks, and tools capable of automatically extracting arguments from natural language. In the area of NLP, there is a wide range of available tools covering various aspects and addressing different challenges. However, there is lack of standardization, accessibility, and acceptability of the existing tools, due to proprietary formats developed for the modeling of natural language in different domains.

The annotation process is of major importance in any NLP system, thus different tools have been proposed following different approaches. The first introduced Web-based open-source annotation tool is BRAT [1] [163], which is based on the STAV text annotation visualizer [164] and is characterized by the wide variety of tasks that can be accomplished and the scientific work that has been conducted using it. It has been adopted in different fields like visualization, entity mention detection, event extraction, coreference resolution, normalization, chunking, dependency syntax, and meta-knowledge annotation.

A tool following the approach of BRAT is WebAnno [38] which has kept the Web interface and visualization capabilities of BRAT and modified the server layer. WebAnno has improved specific weaknesses of BRAT focusing on user and quality management with the addition of monitoring tools and interfaces for crowdsourcing. Currently, WebAnno is in version 3.0 and also offers a Web-instance[2] through the CLARIN-D infrastructure. Both BRAT and WebAnno are an open and live project that could easily be modified in order to include tasks in the sphere of AM, such as argument detection, relation identification or reasoning evaluation.

The construction of graphs for text annotation is followed in GraPAT [158] covering different tasks like sentiment analysis, argumentation structure, rhetorical text structure, and natural visualization of the annotation process. The initial goal for the development of GraPAT was to increase IAA and maximize the automation of annotation without neglecting neither the variability nor the annotation speed/comfort. GraPAT can be considered as the successor of RSTTool [116], as it maintains the principles of Rhetorical Structure Theory

[1] http://brat.nlplab.org/index.html
[2] https://webanno.sfs.uni-tuebingen.de/

(RST) annotation, enriching it with more capabilities like sentiment analysis and argument structure annotation model.

GATE [30] has dominant presence in the wider field of text engineering offering numerous tools and capabilities from simple tasks (e.g. information extraction, named-entity, etc.), to modifications for cutting-edge technologies, such as cloud-enabled software and social media analysis. Regarding the argument annotation task, GATE offers Teamware [16], a Web-based software suite which provides the environment for collaborative annotation and curation. GATE Teamware stands out as the only annotation tool, to the best of our knowledge, which supports execution of an automatic NLP system, before manual annotation.

The DiGAT tool [77] has been developed alongside with an annotation scheme and a graph-based inter-annotator agreement measure based on semantic similarity. Similarly to GraPAT, DiGAT also relies on graph structures for the annotation process, aiming at simple and accurate annotation of relations between entities in long text.

The establishment of the TextCoop platform alongside the Dislog language is presented in [148]. TextCoop is the only tool in this subsection that follows a logic-based approach, heavily influenced by RST, modeling the conclusion as a nucleus and the support as a satellite. As it is described, TextCoop offers a functional Web interface, however, to the best of our knowledge, this is not provided yet. Another tool that is in a similar status with TextCoop is TURKSENT [40], a manual annotation tool with multilingual capabilities aiming at automatic sentiment analysis of text derived from social media. Its mentioned Web-based interface does not seem to be available yet.

Tool	**Web UI**	**Manual Annotation**	**Arg Retrieval**	**Arg Evaluation**
WebAnno [190]	Yes	Yes		
BRAT [163]	Yes	Yes		
GraPAT [158]	Yes	Yes		
DiGAT [77]	Yes	Yes		
MARGOT [97]	Yes		Yes	Yes
OVA+ [69]	Yes	Yes		
TOAST [157]	Yes		Yes	Yes
GATE Temware [16]	Yes	Yes		
Args [180]	Yes		Yes	Yes
ArgumenText [159]	Yes		Yes	Yes
Rationale [176]	Yes	Yes		

Table A.1: A summarization of the existing NLP tools that can enhance the process of AM. The table includes tools in the wider NLP area which can be integrated at any stage of an AM pipeline.

The Stanford CoreNLP toolkit [104] receives great acceptability from the NLP community as it offers a broad range of grammatical analysis tools, different APIs for the majority of the programming languages, and the ability to run as a simple Web service. However, a specific tool for AM-related tasks has not yet been developed. Among the existing tools offered by the toolkit, the Stanford OpenIE is more closely related to AM, as it enables the extraction of relation tuples out of binary relations. Two more tools whchi are included in the toolkit and can be used in an AM pipeline are the Stanford Relation Extractor, which finds relations between two entities located by the Stanford Named Entity Recognizer, and

the Neural Network Dependency Parser, a dependency parser that establishes relationships between "head" words and "modifier" words in a sentence. The Web interface provided by the toolkit[3] offers up to ten different annotators and the visualization of the schemes has been realized using the BRAT software.

The majority of the aforementioned tools are on-going, open-source projects that can be modified in order to carry out AM-related tasks, whereas some others (GraPAT, DiGAT) are graph-based annotation tools that are used for identifying the relations between chunks of text. Other functionalities such as sentiment evaluation and name entity recognition can boost AM related tasks, as sentiment features are used in the majority of the existing research papers in the area as Table 2.1 illustrates. If any of the subtasks that are included in the AM pipeline can be executed automatically and reliably through a software tool, then this tool should be exploited.

A.2 Argument search, retrieval and annotation tools

In this subsection, we present the tools that have been designed to enhance the argumentation process. Some of those tools have been implemented to boost the annotation step, others offer an argumentation search engine, and there are tools capable of automatically grading an argument or even performing the entire process of AM.

The Centre for Argument Technology[4] has produced a series of tools covering different aspects of AM. The latest developed tool is OVA[5] [69] and it has replaced to a certain extent the tool of Araucaria [140]. Arivina[6] [156], the successor of MAgtALO [142], offers a dialogue system implementing the concept of mixed initiative argumentation, where human players and agents debate, having equal levels of participation. The process for the calculation of the acceptability semantics on structured argumentation frameworks is completed through TOAST[7] [157].

Probably the most influential achievement of the considered organization is the establishment of the Argument Web[8] [139], a repository in cooperation with a series of tools, systems and services, such as AIFdb, the main search interface for the Argument Web. ArguBlogging[9] [10] materializes the concept of crowdsourcing in argumentation by capturing arguments that take place in online platforms (tumblr and Blogger are supported) and provides them as feed to the Argument Web. Concerning the educational aspect of argumentation, Argugrader[10] offers automatic grading and provides detailed feedback to students regarding successful or not construction of arguments.

A prototype argument search framework is proposed in [180] which is able to carry out the entire process of an argumentation search engine, from user query and argument retrieval to ranking and presentation of arguments[11]. The argument search engine relies on pre-

[3] http://corenlp.stanford.edu/
[4] http://www.arg-tech.org/
[5] http://ova.arg-tech.org/
[6] http://arvina.arg-tech.org/
[7] http://toast.arg-tech.org/
[8] http://www.argumentinterchange.org/
[9] http://www.argublogging.com/
[10] http://www.argugrader.com/
[11] http://www.arguana.com/args

structured arguments from a defined list of debate portals, and a standard mapping process takes place in order to convert the concepts that characterize each argument in the different debate portals to the common argument model.

A system able to utilize the heterogeneity and big volume data is implemented in Stab et al. [159], under the name ArgumenText. The ArgumenText uses 400 million plain-text documents from different sources and deploys a series of technologies in order to construct a solid pipeline able to materialize a sequence of sub-tasks that eventually present ranked pro and con arguments through a Web interface.

At this moment we are aware of only one tool that accomplishes the complete task of automatic annotation in terms of AM. MARGOT [97] is built on the foundations of Lippi and Torroni [94] and extends the previously established model by including the task of evidence detection and providing a Web interface[12]. The syntactic structures that are followed in argumentative discourse are the fundamental idea on which the tool was built. The model implemented for AM involves two binary classification problems, the argumentative sentence detection and the argument components boundaries detection. The former is addressed with a combination of tree kernel and bag-of-words, whereas for the latter SVM-HMM with bag-of-words, part-of-speech, lemma and named entity are employed. Overall, the tool achieves acceptable evaluation scores, regarding the complexity of the task, but there is still room for improvement in covering various domains.

Another tool that is very similar to MARGOT, sharing the same interface[13], is CLAUDETTE [93], an online platform addressing possible unfair or abusive terms of service. The platform realizes a 2-step algorithm including the binary task of detecting an unfair clause, and if the first step is positive, then a classification task through 8 possible categories takes place. For this classification, a combination of eight SVMs exploiting lexical features is used with fair results in all used metrics.

On industry level, Austhink Pty Ltd. is continuously developing software tools aiming at the improvement of the general reasoning and argumentation process through training sessions. The two most successful software tools are Rationale [176, 151] and bCisive [105], with both of them offering online limited free versions[14][15]. The former supports a series of activities in the area of AM, such as construction, visualization and mapping of arguments, while the latter focuses on providing support to business decision through hypothesis and decision maps. Rationale is considered as the successor of Reason!Able [48].

A synopsis of the existing NLP tools that have been discussed in this section is presented in Table A.1. The existing tools can be classified into three categories, tools that aid the manual annotation process, general-purpose NLP tools and tools that offer an entire mechanism for argument search, retrieval and evaluation. It has to be underlined that the evaluation of the argument differs depending on the different approaches, as in [97] the number of claims and premises are presented, in [157] the weight of the argument is calculated and in [180, 159] the arguments are categorized as pro and con.

[12] http://margot.disi.unibo.it/
[13] http://155.185.228.137/claudette/
[14] https://www.rationaleonline.com/editor/
[15] https://www.bcisiveonline.com/editor/