Improving Problem-Oriented Policing with Natural Language Processing

Anthony Dixon

Submitted in accordance with the requirements for the degree of Doctor of Philosophy.

The University of Leeds, School of Law

May 2023

Published Work

Dixon, A., & Birks, D. (2021). Improving policing with natural language processing. In Proceedings of the 1st workshop on NLP for Positive Impact (pp. 115-124).

The above work was completed with my supervisor Daniel Birks and is based on the material covered in Part 1. I completed the research for the article and drafted it. Dr Daniel Birks mentored me through the process and assisted with editing.

Halford, E., Dixon, A., & Farrell, G. (2022). Anti-social behaviour in the coronavirus pandemic. Crime Science, 11(1).

This work relies upon the analysis that was conducted in study 2. A full explanation of the relationship to this work is given at the start of study 2.

Acknowledgements

I would like to thank my supervisors Dr Daniel Birks, Prof Graham Farrell and Prof Nick Malleson for their support throughout my research. I'm particularly grateful to Dr Dan Birks for his work to set-up the research and for his ongoing support throughout my research. I am also grateful to my supervisors for inviting me to join them as part of the "Covid and Crime harms" research team. This research opportunity was a wonderful experience and truly defined my time as a post-graduate researcher.

For a large part of my time during this research I worked from home, largely due to the pandemic. As a result, my co-workers were my children Lucy, Stanley and George. I thank them for being wonderfully enjoyable co-workers. I will look back at this time, and the memories we created together, with real joy. I feel enormously lucky to have had the opportunity to have spent that extra time with them. Thanks also go to my parents, who continue to demonstrate that no matter how old you are - your parents can always make sure you stay on track!

Finally, the biggest thanks go to my wife Katie. Katie supported me throughout my career in the Army, sacrificing her career as we moved around the country (and the globe). So, to my surprise, and eternal gratitude, she did not even raise an eyebrow when I suggested that I extend my agreed one-year employment hiatus to four in order to complete a PhD. None of this would have been possible without her support. Thank you Katie.

This work was supported by the Economic and Social Research Council grant number ES/P000746/1.

Abstract

The policing approach known as Problem oriented policing (POP) was outlined by Herman Goldstein in 1979. Despite POP being shown as an effective method to reduce crime it is difficult to implement because of the high analytical burden that accompanies it. This analytical burden is centred on understanding the mechanism by which a crime took place. One of the factors that contributes to this high burden is that a lot of the required information is stored in free-text data, which has traditionally not been in a format suitable for aggregate analysis. However, advances in machine learning, in particular natural language processing, are lowering the barriers for extracting information from free-text data.

This thesis explores the potential for pre-trained language models (PTMs) to efficiently unlock the information in police crime free-text data. PTMs are a new class of machine learning model that are 'pre-trained' to recognise the meaning of language. This allows the PTM to interrogate large quantities of free-text data. Thanks to this pre-training, PTMs can be adapted to specific natural language processing tasks with much less effort. Efficiently unlocking the information in the police free-text crime data should reduce the analytical burden for POP. In turn, the lower analytical burden should facilitate the wider adoption of POP. The thesis concludes that the evidence suggests PTMs are potentially an efficient method for extracting useful information from police free text data.

Contents

List of Figures

List of Tables

Part I

Background

Chapter 1

Introduction

1.1 Motivation

This research is located at the intersection of a well-established crime reduction methodology, problem-oriented policing (POP), and a growing field in artificial intelligence, natural language processing (NLP), which is increasingly making it easier to draw information from unstructured data.

POP is a method of policing first introduced in 1979 by Herman Goldstein (Goldstein, 1979) . POP replaces the traditional policing model, which focusses on responding to single incidents as they occur. By contrast, POP seeks to prevent problems (problems are defined below but are essentially any issue the police are expected to deal with) from reoccurring by analysing how they occurred in the first place and then intervening in the generation process. In this regard, an essential element for conducting POP is understanding the conditions that allowed the problem to occur. Crimes are a subset of the problems that police forces face, albeit a large and important one. POP's focus is on problems, not just crimes. However, the POP terminology sometimes focusses on crimes only. Where the terminology does so, this is normally without loss of generality of the effect of POP on all problems encountered by the police.

POP seeks to tackle problems, which it defines as a "A cluster of similar incidents, whether crimes or acts of disorder, that the police are expected to handle" (Scott, Eck, Johannes, & Goldstein, 2016). Of immediate interest from this definition is that problems should have similar incidents, which is related not just to the outcomes but also to the processes and external factors that occur leading up to, during and after the incident. As an example if a problem is burglaries in an area, then an incident would be the individual burglaries. Most of the analytical effort required in POP is expended in scanning for and then grouping similar crimes, followed by analysing incidents to identify similar factors influencing the incident occurrence.

This intersection between POP and NLP is important. Although POP works (Hinkle, Weisburd, Telep, & Petersen, 2020), it is necessary, yet seemingly difficult, to follow the POP framework correctly. Thus, although POP has shown benefits, it has not realised its full potential (Sidebottom, Bullock, et al., 2020). The impediment to POP that this body of research aims to reduce is the analytical burden necessary to understand the specificity of the problem or problems at hand. POP works best by attacking the mechanisms of the problems so that the opportunities to commit the crime (or other non-criminal activity) are significantly reduced (Clarke & Eck, 2003).This is achieved by understanding the causes and mechanism of the problems and then finding ameliorating strategies. Understanding problems so that they can be attacked and grouping problems so that solutions can be used efficiently are key components of POP. However, studies have shown that the analytical power and data required to do this efficiently are difficult for police forces to muster and coordinate (Sidebottom, Kirby, et al., 2020). This research will show how NLP models can efficiently extract the required information to enable POP interventions.

Although police forces have a mandate to record all crime (Home Office, 2020), the bulk of the recorded information about crime is contained in textual data, such as in police generated crime notes, witness statements or forensic reports. Accessing this information is largely completed manually (Goldstein, 1990), and as such, it is often a long and laborious task. Given the resource pressures, the work must be completed selectively (Rogerson, 2016). Unlocking access to this information would enable analysts and officers practical access to a much wider source of information with which to do their job. Two sources of text data are introduced next as potential sources of information.

MO data are relatively short sections of text of around three to eleven sentences that describe what is initially known about the crime. The text is generally limited to the knowledge that can be gathered by the initial responding officers from their provisional review of the crime scene and any victim or witness statements. Further investigations, for instance by detectives or forensics, are held in the case reports and are not detailed in the MO data. As such, they offer a concise but limited view of the crime. Alongside the MO data, more typical crime data is recorded in a structured way, with fields such as time, date, location, crime classification and victim characteristics often included. These more structured data have been exploited to a greater extent than the unstructured MO data. See Braga, Papachristos, and Hureau (2014), Johnson (2016), Ratcliffe and McCullagh (1998), Weisel (2016) for a selection of methods.

A second source used in this thesis is police incident logs. These are usually generated by a call operator who responds to emergency and nonemergency calls from the public. Typically, they record the details of an incident as it is in progress. More recently, incident logs have also started to encompass reports from the public that have been logged electronically, through email or online forms. Police incident logs differ from MO data in two important respects. First, they are not generally edited, in that they provide multiple perspectives over time rather than a single post-hoc view. Second, they cover both crime and non-crime incidents, so they have a much broader reach than MO data. The two data types are discussed more extensively in Chapter 8.

Recent advances in NLP (NLP is a sub-set of Machine Learning (ML)), where the basis of models has moved from a more logical and rules-based approach to a more probabilistic approach, have allowed more powerful models to be applied to free-text problems (Kumar, 2011). Improvements in processing power and the availability of data have also pushed the boundaries of the state-of-the-art (SOTA) models. The improvements in NLP have led to the development of suites of generic open-source tools (Benoit et al., 2018; Loper & Bird, 2002; Manning et al., 2014). These toolkits are designed so that they can be reused on different sets of natural language texts to solve similar problems, such as classification or question and answering, without the need to build models from scratch each time a problem is encountered.

Pre-trained language models (PTMs) are an import class of these generic NLP tools. PTMs are different to the normal classes of models because they are built in two parts. The first part of the training is relatively generic and trains the model to “understand” language. The second stage trains the model on a specific task. That is PTMs are firstly *trained* to understand a language eg. English before they are *fine-tuned* on a specific task e.g. classifying burglary MOs. PTMs will be explained more fully in the NLP chapter.

This research focuses on investigating the use of PTMs with police free text data, in particular classifying short police texts to identify intra-crime variation. While the PTMs have been found to work well in other domains, they have yet to be tested on free text generated by the police on problems that are important to the police. Can these PTMs be leveraged by the police and therefore taken advantage of for their lower barriers to use? That is the fundamental question of this thesis.

1.2 Research Questions

The research question and its supporting objectives are stated here with a brief explanation to guide the reader through the next few sections. The research questions are examined more thoroughly in relation to the literature outlined in the rest of the document in Chapter 7. The main research question is as follows:

Can PTMs be used efficiently to extract information from police free-text data, and if so what practical applications for problem-oriented policing does this approach have?

In this thesis extracting information will focus on automatically classifying texts to understand if an event did or did not happen. For example Burglary MO texts will be classified to understand if the burglar used force to enter a property or not.

The research supporting objectives are:

1. **Identify the extent of NLP usage with police data.** This is largely conducted in the literature survey, which is the focus of Chapter 6.
2. **Evaluate how effective PTMs are with MO data.** PTMs are formally introduced and explained in Chapter 5. MO data is introduced in Chapter 8. Study 1 investigates the use of PTMs to classify MO text data.

3. **Evaluate how effective PTMs are with Police Incident data.** As mentioned above, police incident data is another source of information on problems the police face. Using PTMs to classify police incident logs is investigated in Study 2.

4. **Evaluate how effective Active Learning is with police data.** Active learning is a method to reduce the amount of data that PTMs need to learn. It has been found to work with other data types, but its effectiveness with police data is unknown. Active learning is introduced in Chapter 4 and studied in study 1.

5. **Identify which parts of the POP process might be best supported by the use of PTMs.** The POP process is explained fully in Chapter 3. It is likely that different parts of the process will find differing uses and utility for PTMs. Lessons for POP are drawn from both studies and outlined in Part 3.

6. **Identify implementation barriers for PTMs.** Any new process is likely to have implementation barriers, which are important to identify so that they can be minimised. Discussed in Part 3.

1.3 Thesis Structure

This thesis has three parts. Part 1 focusses on the introduction and background to the research. Part 2 presents studies exploring the use of NLP with police free text data. Part 3 draws on the research of Part 2 and explores the implications for POP.

Part 1 begins with an introduction, which sets out the main ideas of the research. The next chapters then explain the theoretical underpinnings of POP, namely routine activity theory and situational crime theory. POP is then discussed in more detail, identifying key problems with widespread usage. After POP is explained, the focus switches to the more technical aspects of the research. First, machine learning (ML) is introduced. Next, ML with free-text data, namely NLP, is explored including the theory and use behind PTMs. The penultimate chapter draws these POP and NLP together by conducting a literature survey of the use of NLP with police generated free-text data.

Part 2 focuses on the two main study areas, which are delineated by the type of police free text data used. Both studies focus on the utility of PTMs with police generated free text. Study 1 uses MO data. Study 2 uses police incident log data. Study 1 is split into three parts. The first part, study 1a, investigates the classification of MO texts in one police force area (known as PF1). Study 1b investigates the efficiency of active learning, using the data and models from study 1a. Study 1c replicates and extends study 1a using data from a separate police force (PF2). Study 2 only has one part and is focussed on classifying police incident logs in a single police force (PF2). There is a table at the end of Chapter 7 (7.1 that captures this detail).

The final section of the thesis, Part 3, summarises the lessons and implications from the studies in Part 2 in light of the conclusions from Part 1. It does so in two chapters. The first chapter discusses the implication of NLP usage for POP, particularly how and where PTMs might be used to alleviate the analytical burden. The second chapter takes a broader look at potential future research directions of PTMs with police free text data.

Chapter 2

Core Related Theoretical Frameworks

This section introduces some especially relevant theories form the wider research field of crime science that underpin the general approach of POP. The routine activity theory (Cohen & Felson, 1979)and then situational crime prevention (Clarke, 1997) are explored individually to help build the concepts on which POP is based.

2.1 Routine Activity Theory

First introduced by Cohen and Felson in 1979, Routine Activity Theory was proposed as a theory to help explain the increase in crime after WW2. The change that the theory brought about was a shift from thinking about crime purely as a social process to seeing it more as a socio-physical world (Felson, 2016). The focus was on the crime event itself and what conditions were needed for the event to be created. This focus on the crime event has obvious parallels with the focus on the problem in POP, and indeed, the theory has been extended since its first inception to move beyond crime.

In their original article, (Cohen & Felson, 1979), Felson and Cohen sought to explain the crime event through the convergence of three principle physical aspects, that is 1) a likely offender, 2) a suitable target and 3) the absence of a

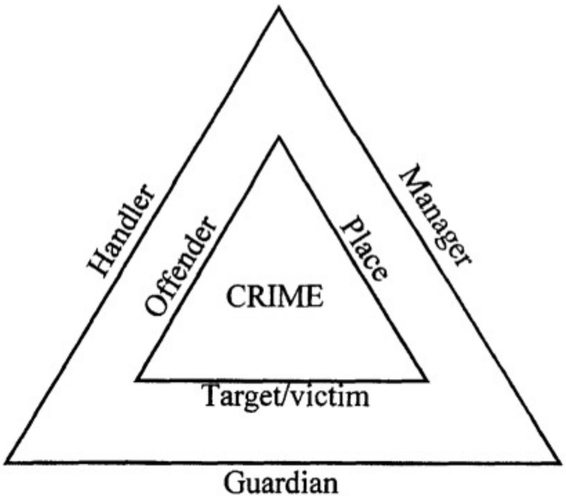


Figure 2.1: Eck's Crime Triangle, reproduced from Eck, 2003

capable guardian. These three elements, modified and extended by (Eck, 2003) to form the problem triangle in Figure 2.1, demonstrate the close ties between the two bodies of work. For instance, if a crime or problem opportunity is generated through people's movements and the types of activities they conduct, then by extension, modifying these activities should also affect the prevalence of crime opportunities.

Therefore, from a POP standpoint, thanks to routine activities theory, there are now at least three broad opportunities to prevent crime. That is, by adapting one of those three physical aspects outlined above, the problem triangle can be broken, and the problem opportunity is lost. This contrasts with the traditional model of policing, which concentrates on a narrow aspect of the offender, namely dissuading him or her through a deterrent effect (police response is likely to catch you) or a removal effect (locking them up prevents their ability to commit crime outside of confinement).

The problem triangle in Figure 2.1, with its added outer layer, suggests three broad means with which it can be broken to eradicate the opportunity. The handler has an effect on the potential offender – perhaps their presence physically or emotionally makes the offender less likely to commit an untoward act. The guardian protects the target from would be offenders, and this can vary from a person actively guarding their luggage to unintentional increased footfall in residential areas reducing the opportunity for burglary (Halford, Dixon, Farrell, Malleson, & Tilley, 2020). Place managers govern how a place functions (Scott et al., 2016). They may be bar managers, shop designers or teachers. They play an important part in the opportunity structures that arise through the way business is conducted and how the physical environment is set out. Identifying problems and influencing this group of people to change their environment is a good example of a strategy originating from the POP framework.

2.2 Situational Crime Prevention

Situation crime prevention (SCP) rests on this claim "Reducing Opportunities for specific forms of crime will reduce the overall amounts of crime"(Clarke, 2016). Like POP, this theory focusses on the target and the place of crime – "It seeks to forestall the occurrence of crime, rather than to detect and sanction offenders"(Clarke, 1997). focussing on what can be changed now to have an almost immediate effect on the cause of crimes (Clarke, 1995). The principles for situational crime prevention are very similar to POP, as explored below:

1. **Focus on specific categories of crime.** Situational crime prevention works best by only attempting to tackle one type of crime at a time, calling for specificity in defining how these crimes are conducted and how the opportunities have been generated (Felson & Clarke, 1998).If the categories are grouped too widely in the first instance, then common patterns will not be found, and common solutions will be unlikely to work.

2. **Understand how the crime is committed.** The focus is on how, not why the crimes were committed – by understanding how they were committed, the mechanism can be interrupted, and the crime can be prevented. Again, any information that is related to how a crime is committed will be useful in its prevention. It is important to note that while some of this information may be found in police reports, they are unlikely to reflect the full range of actions before and after the criminal act.

3. **Use an Action Research model.** "Action research is an iterative process involving researchers and practitioners acting together on

a particular cycle of activities, including problem diagnosis, action intervention, and reflective learning." (Avison, Lau, Myers, & Nielsen, 1999). To most practitioners, this probably means doing what you normally do, *observing* the problem, *orientating* to the problem, *deciding* what to do and *acting* on that information, correcting as one proceeds by continually cycling through these stages (the OODA loop as developed by John Boyyd (Gray, 1999)). However, it may be useful to highlight the importance of both deeply understanding a problem and acting on that information to combat the crime, rather than using the two strands in isolation.

4. **Consider a variety of solutions.** Be open to a whole host of solutions in order to bring around the desired effect. A selection (twenty-five) of generic solutions have been posited as a good starter from which to initially pick and then adapt solutions to implement. Although too many to list here the five groupings of situational crime prevention give a flavour of the spread of solutions that are available. The groupings are as follows:

    (a) **Increase the effort.** Make crimes harder to commit such as an additional layer of security to overcome e.g. a steering lock on a car.

    (b) **Increase the risks.** Generally increase the chance of being caught, such as through increased surveillance.

    (c) **Reduce the rewards.** Make the crime less attractive. For example, by making stolen goods harder to sell, their value is decreased, and the rewards reduced.

    (d) **Reduce the provocation.** Lessen flash points. An example might be to make bars less crowded to reduced unwelcome interactions.

    (e) **Remove excuses.** Provide obvious information so that ignorance can not be used as an excuse. Erect signs to remind potential offenders of rules in specific areas.

Situational crime prevention has been criticised on a number of levels for being a superficial technique for reducing crime (Wortley, 2010). Some of the more relevant critiques to this research include:

1. **Crime is not reduced only displaced.** Some critics, especially those who believe that the amount of crime is largely driven by people's

propensity to commit crime and not by situational factors, believe that preventing crime in one area will result in a similar increase in crime in another area – that is, the crime will be displaced rather than prevented. However, a systematic review into this issue (Guerette & Bowers, 2009) found that while there are some instances of crime moving to other areas after an intervention, the net benefit was still a reduction in crime. However, they do indicate that controlling in different areas for displacement is difficult, as the displacement may manifest itself temporally, spatially or even to different crimes altogether.Guerette and Bowers (2009) also do not seem to control for publication bias, which may have a detrimental effect on their results.

2. **SCP doesn't work for expressive crimes.** Although critics accept that high-volume acquisitive crimes can be reduced, they believe that crimes that are more expressive or are *irrational* will not be as easily affected by situational crime prevention techniques. Expressive crimes include domestic violence, sexual offences or those committed in the heat of the moment. However, within the toolkit of SCP, there are efforts to reduce provocation that may prevent manifestation of expressive crime conditions, and there are examples of situational crime prevention projects hosted on the virtual problem orientated policing centre. However, given that (Guerette & Bowers, 2009) do not explicitly mention expressive crimes for the studies they used for the systematic review, there is perhaps a practical deficit, if not a theoretical one.

2.3 Conclusion

This section has demonstrated that crimes can be prevented from occurring by disrupting the process that create conditions for that crime. The processes centre around the target of the crime, the place that the target is in and the offender. Disrupting the conditions of any these three factors can be enough to prevent the occurrence of a crime, in much the same way as preventing the coming together of the fire triangle elements prevents a fire. Situational crime prevention has been developed to exploit this principal and by doing so has posited five main intervention types that can be used to reduce crime. However, for these interventions to work, they need to be aligned to the crime and its context. The next chapter builds on these theories by introducing POP in more detail.

Chapter 3

Problem Oriented Policing

As described previously, POP is a policing model introduced in 1979 to ameliorate some of the shortcomings stemming from the traditional policing model of response. The general aim is to tackle the factors that allow a crime opportunity to occur so that it can no longer occur in the future. In this way, the overall aim is crime (or problem) prevention. The next section explains in more detail the core principles of POP. This is followed by a thorough exposition of SARA, an analytical framework for the conduct of POP, to demonstrate how it is conducted in practice. Finally, there is an evaluation of the utility of POP, before assessing POPs weaknesses.

3.1 Overview of POP

Problem-oriented policing was introduced by Herman Goldstein in 1979 (Goldstein, 1979) as a new policing model to replace the traditional response policing style. Since its introduction, POP has been widely utilised as a method to reduce problems faced by police services across the world (*Fairness and Effectiveness in Policing: The Evidence*, 2004; Goldstein, 2018).The central thrust of POP is to focus on the ends of police activity (e.g., reductions in harm to the public) rather than the means (e.g., number of convictions). Recognising that the police have a wide variety of objectives to deal with, the focus should be on the resolution of these problems, not on the means or the ways of addressing these problems.

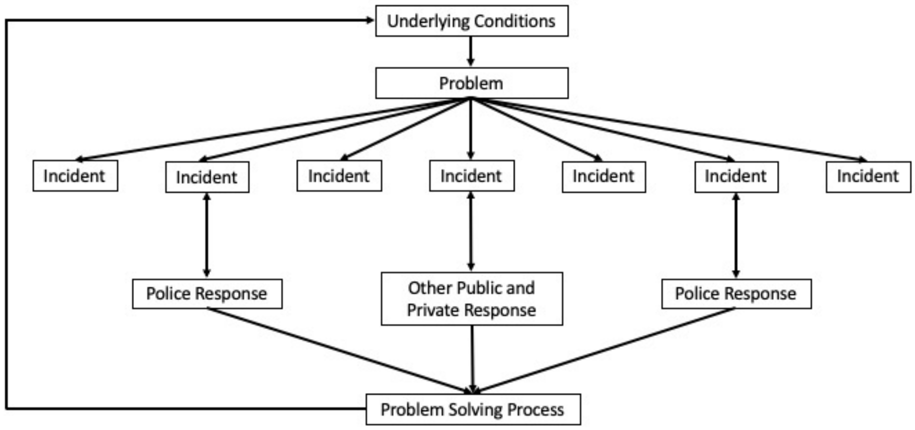


Figure 3.1: A schematic for POP. Reproduced from Eck and Spelman (1987)

The mechanism for this prevention is demonstrated in the schematic at Figure 3.1. The incidents are generated from an underlying causes; however, in the traditional response model, the incidents are responded to individually, and, typically, due to resource constraints, not all incidents are known about or can be resolved. The POP model takes the information from these responses and the similar incidents and aims to tackle the underlying causes so that the opportunity to create the problem no longer exists (Eck & Spelman, 1987). In broad strokes, this is how POP aims to reduce harms, by using knowledge of similar incidents and then altering the crime triangle so that the conditions are no longer present for the problem to occur.

Police business is not just about crime, it is about all problems that the police are responsible for or are thought to be responsible for. Problems can be defined as “Problems are a cluster of harmful incidents that the public expects the police to handle” (Scott et al., 2016). However, these problems should have a common theme, so that they can be grouped and addressed together. Although most of the focus may be on crimes, POP does not exclusively focus on crime and recognises that the police remit is much wider than crime alone.

The next few sections introduce and explain some of the core principles surrounding POP. When considering these principles, it is important to keep in mind that “POP is a framework, or methodology, for addressing police problems and not an intervention strategy per se” (Scott & Clarke, 2020). That is, not all of these principles are directly required for the solving of “a” problem, but they are required to build a culture of problem solving in general.

3.1.1 Focus on Harm Reduction

POP places more emphasis on preventative responses rather than remedial ones, enforcing the age-old heuristic that “prevention is better than cure”. Remedial action – acts of immediate response, investigations and arrests – is the staple diet for the traditional model of policing, and POP seeks to move away from these defaults and to act before the crime or problem arises. This does not mean that these elements do not have their place; rather, the emphasis is on rebalancing the focus between remedial and preventative action (Goldstein, 1990).

This focus on harm prevention switches the measure of effectiveness of the police, moving away from more managerial approaches of clear-up rates and arrest statistics to a more considered view of the police’s effectiveness around reducing problems (Goldstein, 1990). Eck and Spelman (1987) suggest that there are several ways that effectiveness in POP can be measured, and that it does not rest solely on problem elimination, but also on the reduction of similar incidents, the seriousness of those incidents and how they are responded to. The focus is squarely on the incidents and the characteristics of the future occurrences, such as frequency and severity. The focus is not on the more traditional metrics of police success, such as arrests or response times.

The focus of POP is not on catching criminals but preventing problems. If problem opportunities are not presented, incidents cannot occur in the first place, and criminals cannot cause harm to the population to be protected, which situational crime theory shows to be a real possibility that must be explored. However, as Figure 3.1 shows, grouping of incidents with the same underlying cause is necessary for the efficiency of reducing many incidents from a single intervention. This is the focus of the next principle.

3.1.2 Specificity! Specificity! My Kingdom for some specificity! [1]

As shown above, the central idea behind POP is to reduce incidents of harm by disrupting the underlying causes of the problem. This disruption is achieved through the grouping of individual incidents into problems, understanding the similar mechanisms that cause the problems and disrupting these mechanisms so that the problem can no longer occur.

Identifying these groups or clusters of problems can be difficult and as Scott and Kirby (2012) highlight, there is a tension between breadth and depth of knowledge of problems in organisations. In short, the higher one goes up in the organisational hierarchy, the wider one can see problem occurrences. Further, one gains greater breadth of the situation, and therefore the efficiency of POP can increase as more single incidents can be grouped. However, this increased breadth comes at the expense of the detailed knowledge of each problem that can be found lower in the hierarchical order, allowing each group of problems to enjoy greater intrasimilarity. This tension between breadth and depth of problem knowledge can inhibit the optimal implementation of preventative measures (Maguire, Uchida, & Hassell, 2015).

Once the incidents have been grouped, it is necessary to understand them separately. This examination is not necessarily about individual elements, but about the similarities of mechanisms between the individual acts that make up the problem set. The focus is more on the “why” than the “what” or “who” of traditional policing. Why did this problem arise? How did the circumstances around each problem set the conditions for the harmful act to occur? What are the common factors between problems? Answering these questions with fine-grained analysis leads to a deeper understanding of the problem itself. Understanding the steps and conditions that lead to the problem means that points can be identified and tackled to prevent the conditions for the harmful act being realised, reflecting routine activity theory and the principles behind situational crime prevention (Felson & Clarke, 1998).

As Felson and Clarke (1998) highlight, “crime opportunities are highly specific”, that is, they should be understood and grouped by how they have been committed and not necessarily by what the outcome of the problem was. The

[1] With apologies to Shakespeare’s *Richard III.*

key element for POP, therefore, is specificity in fitting a solution to a well-developed problem. Specificity is both the Achilles heel and the Herculean strength (apologies to readers for the mixing of ancient metaphors) of POP. So, although POP is effective, it is also difficult to achieve. Where effort in specifying the problem falls short, this directly influences the effectiveness of the solution and hence the final result (Maguire et al., 2015). Therefore, any effort to make the analysis of a problem easier, more effective or more efficient will have a disproportionate effect in the success of POP.

3.1.3 Tailored Responses

The microscopic evaluation of the problem allows a new approach to be taken for each problem. Each problem will undoubtedly have its own set of conditions and unique factors, and by understanding these a new and problem specific strategy can be developed to tackle that particular problem. This is the main thrust of the approach. Pick a solution that is effective for the problem set at hand, which is achieved through understanding the problem thoroughly.

Set against a back drop of the traditional policing model of responding to crime incidents, POP sought to expand the repertoire of police responses by encouraging the use of tools other than the criminal justice system. Criminal justice systems can be slow and inefficient, and may not do a good job of ameliorating the harm that has occurred. With a focus on prevention it is necessary to look outside the traditional toolbox of police responses to find a new set of tools. This new set of tools will allow the leverage of other capabilities in the public and private sectors that can be utilised to change the conditions that allow problems to flourish. Reflecting on POP in 2018, Goldstein (2018), reflects on the success and “enormous potential” of the use of non-police entities to reduce crime by using their powers or resources.

Formulating tailored responses is resource intensive as the problems need to be extensively detailed and a fitting solution found. In order to make the formulation of the response less onerous there is a heavy emphasis on reporting and logging results so that inspiration, though not exact solutions, can be used to formulate tailored responses.

3.1.4 Evaluate the results

With a rigorous focus on reducing harm, it is important that POP has within its framework an emphasis on evaluating how well it is achieving its stated aim. There needs to be an understanding of which POP implementations have worked, which have not and why. This helps not only to ensure that the actions are having the desired consequences, but also to make the implementation of the process more efficient, by building a body of knowledge that can be used by all practitioners. Proving that something has not happened, a counter-factual, is always more difficult than demonstrating an occurrence. Additionally, attributing that non-occurrence to a specific intervention can be even harder. That is why the evidence for the utility (positive or negative) of POP must be actively sought. The measurement must begin at the outset and may even need to encompass an area much wider than the target zone. Measurement will be difficult to achieve in retrospect alone. Identifying weak signals in noisy environments is difficult, so the use of analytical techniques that are not routinely found in the police organisations is likely to be necessary, thus adding to the analytical burden (Scott et al., 2016).

In addition to understanding internally whether a POP intervention has been effective, it is also necessary to publicise the results. As shown above, the introduction of POP is a change to the norm. It is not the de facto style, and it is not what most officers in police forces envisioned they would be doing when they joined the organisation. Reporting the results is crucial to building an understanding of whether POP works, and therefore is a worthwhile activity for the police to engage in.

The results reported will help to build a body of knowledge about what works. Given the focus on specificity of problems, solutions are unlikely to be ported wholesale from one area to the next. However, building a body of knowledge is important for two reasons. First, it will allow some of the analytical burden for each round of problem solving to be completed more quickly, as drawing on the experiences of others will allow adaptations of plans or a swifter understanding of mechanisms that can then be adapted. Second, the body of knowledge will act as a beacon for the effectiveness of POP and a fulcrum for the turning of the tide of institutional resistance.

3.1.5 POP Summary

Returning to Figure 3.1 the essence of POP is about changing the underlying conditions that allow crimes or problems to flourish. These changes are brought about by applying analytical power to first group problems and then analyse their structure to find a suitably specific response. Once this response has been implemented, there is further need to document the effect and report the results to contribute to a wider body of knowledge to develop to the understanding and efficient conduct of POP. This section was about what POP is. The next section will take a deeper look at how POP is conducted.

3.2 SARA - An Analytical Framework for POP

Although POP can be implemented by police forces in several ways Scott and Kirby (2012) suggests there are two broad implementations of POP in a police force: either having all officers conduct POP, the generalist approach, or building specific capability and units, using a more specialist approach. No matter how POP is implemented, the broad analytical process follows tends to be centred on what is known as the SARA process (Sidebottom, Bullock, et al., 2020).

SARA stands for Scan-Analyse-Respond-Assess and the cycle is shown in Figure 3.2. Clearly, the type of implementation for POP depends on the depth to which the SARA process can be used. However, there is a general flexibility within the model to account for those differences. The POP guide "Become a problem-solving crime analyst in 55 steps" (Clarke & Eck, 2003) is a key document for the implementation of POP in the UK, and sets out how the SARA process should be followed. What follows is a brief look at the four stages of SARA, as described in "55 Steps" and how they interact to form the lifecycle of a problem solving process.

3.2.1 Scan For Problems

The first stage in the process is to scan for a problem, and here it is worth remembering exactly what a problem is. In his book Goldstein (1990) Goldstein

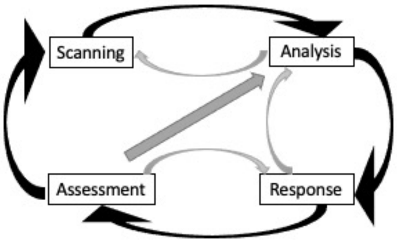


Figure 3.2: The SARA problem-solving process. Source Clarke and Eck (2003)

defines a problem as:

1. A cluster of similar, related, or recurring incidents rather than a single incident.
2. A substantive community concern.
3. A unit of police business.

This definition is quite broad, but it does allow for an understanding to be formed about what one should be looking for when searching for problems. Particularly, the problem must be reoccurring and have a negative effect on the community. The type of POP implementation in an organisation (generalist or specialist) depends on what scanning horizons will be used. If POP is disaggregated throughout the force (generalist), many sensors will pick up on smaller collections of problems but in finer detail. Where POP is more centralised (specialist), the view of the POP scan will be much wider, but will suffer from a lack of detail, because either the information required is recorded but is hard to access, or it is simply in the heads of the officers closest to the problem (Goldstein, 1990). Such trade-offs are inevitable in large organisations, but being able to either widen a scanning horizon or detail more information about each problem is likely to move closer to lessening the severity of the trade-off required.

The scan in “55 steps” is focussed heavily on defining the problems once they have been found in the scan phase - but this presumes that the problems have already be identified. If the problem is a unit of police business then the sub-element has to be a single incident (as seen in Figure 3.1). The scanning phase identifies these incidents and characterises them to group them into a single problem. Then, the job of more clearly defining the boundaries of the problem can begin in the next stage. It is important to note at this stage that although problems are focussed on harms to the community, Maguire et al. (2015) highlights that most of the routes through which cases are nominated for POP action involve police data (70%), meaning that extracting and stratifying police data is likely to lead to improvements to problem identification and formulation.

3.2.2 Analyse in Depth

The problem has been selected, and its boundaries have been broadly defined. Now, it is time to fully understand the problem to refine development. It is at this stage that specific details of the problem are developed, which sets it apart from others and lays bare the underpinning processes that generate the opportunity for the problem to exist.

This stage requires all aspects of the problem and its incidents to be understood(Clarke & Eck, 2003). This will broadly involve trying to understand all the actors involved in the problem, which include the more obvious examples of victims and offenders but also other actors that might be identified from the problem triangle, such as offender handlers and place managers. In addition to understanding the actors, knowing the contexts of the incidents, including any important physical or social factors that led up to or resulted in the problem, will help to identify similarities and pinch points where preventions can be directed.

To acquire the information to characterise the problem, POP practitioners should consider a variety of information sources, which should include the established literature. The POP centre [2] has a wide range of literature that includes specific problem guides as well as academic articles on POP successes. Police files, which include the full gambit of documents including witness

[2] https://popcenter.asu.edu

statements and forensic reports, will be a vital source of information, though they do have their drawbacks, as they often do not reflect the whole problem process. In addition to these written sources, speaking with the original police officers that dealt with the incidents, the victims, witnesses and the offenders can be a rich source of information to enable problem understanding (Goldstein, 1990).

Understanding problems at this level of detail requires a concerted analytical effort, which cannot easily be found in a police force that is not geared towards an analytical approach. This analytical burden is reflected in the results of a review into POP for England and Wales (Sidebottom, Bullock, et al., 2020), where results showed that analysis in POP investigations frequently only included one type of data, rather than the variety highlighted above. Most investigations barely moved above a cursory exposition of simple crime count data. To further highlight the problem, over half of the respondents said they lacked enough analysts to complete this phase properly (Sidebottom, Bullock, et al., 2020).

3.2.3 Respond

Once the problems have been found and analysed the conditions allowing for problems to occur should be clear. These conditions can now be disrupted with a response.

Find a practical response is how it is framed in “55 Steps” and they draw heavily on the five methods of situational crime prevention - mentioned above - to begin to systematically investigate responses. Of note here is the POP centre webpage [3] which includes 74 problem solving guides, highlighting those responses that have been used before and found to work. The responses need to be appropriate and need to be aligned with the work found in the analysis stage. POP puts an emphasis on non-enforcement activity and preventative measures. These potential measures are not limited to those actions that the police can conduct themselves but are part of a wider community approach.

[3] https://popcenter.asu.edu/all-problems

3.2.4 Assess

As set out earlier, assessing the effectiveness of responses employed is beneficial both for proving that POP works and for detecting success, which may otherwise not be as obvious as traditional methods. It is also important to look at any diffusion of benefits that may have occurred around the target area or within the target area but of a different nature. Again, these metrics can be hard to grip, and comparing them to other areas may be necessary to highlight differences to the counterfactual situation where the intervention did not occur. Noisy data may make it especially difficult to pick out weak signals, and changes may not fall across existing recording criteria. All these factors mean that thorough problem and technical knowledge will be required before, during and after an intervention to ensure that the full impact of the intervention is known.

The SARA framework is a cycle, and as practitioners come to this point in the cycle, they should begin again from the beginning – building on their knowledge of the problem by further refining the details and then any additional required responses will make best use of the analytical framework.

3.3 Success of POP

POP is generally regarded as a successful method. There have been two Campbell systematic reviews into the effectiveness of POP, and both have shown that, overall, POP is successful in reducing the problems it has set out to tackle.

The first review, (Weisburd, Telep, Hinkle, & Eck, 2010), showed a moderate indication of success. Although only a strict meta-analysis was conducted against ten studies, they found a small but positive effect in favour of POP. The research papers did not compare directly against other policing models such as intelligence led or community policing.

Additionally, because the reviewers found so few studies that met the criteria for the systematic Campbell review, they also reviewed 'before and after' studies that did not meet the full criteria. In reviewing the additional forty-five "before and after" studies, they found reductions in problems of up to 35%; that is, if the standard model allowed 100 problem incidents to happen in a unit time, then after a POP intervention, this may have been reduced to between 70 and 80 incidents. Of course, what this does not account for is the net benefit of less problems, and this would be hard to quantify; however, if less incidents are responded to, then more police resources are available for other tasks (perhaps even more crime prevention). Further, if hypotheses such as the debut crime hypothesis and the keystone crime hypothesis (Farrell, Laycock, & Tilley, 2015) are true then the benefits over the longer term for some problems will almost certainly be larger as the effects compound.

The second, updated, review, (Hinkle et al., 2020), was able to include many more studies in the main analysis (34), as the quality of the formal evaluation process has increased in the ten-year interval between the two studies. This second review has found an even larger effect, using the stricter criteria, and even managed to quantify the benefits in diffusion with no crime displacement from POP activities. That is, areas surrounding the POP interventions generally also saw a net decrease in those problems (known as a diffusion of benefits).

Another review, this time into hot-spot policing, (Braga et al., 2014), also found a further reduction in crime when problem solving techniques were used alongside hot-spot techniques. As a control, the researchers used studies that employed hotspot policing coupled with a more traditional approach.

We have demonstrated here that POP can be successful, and indeed generally is. But what is represented by these studies amounts to an analysis on a "per-protocol basis" where only those POP interventions that followed the protocol (SARA) were measured in the study, and in practice a more thorough picture of the merits of POP would be based on an "intention-to-treat basis" that would highlight where POP could of been used or was partially used. This would add additional knowledge around what works, but also what does not work or what is impeding POP. What is known about impediments to POP is discussed in the next section.

3.4 Impediments to POP

Although POP has been shown to be effective in reducing crime, the fact that only 34 effective surveys in 40 years of policing were available for the second Campbell review can be seen as evidence of a lack of widespread and sustained adoption. In fact, many studies and reviews of POP have found that although POP reduces crime, it is difficult to implement. In an accompanying article to the first Campbell review mentioned above, Tilley (2010) cites three reasons for POP not working as well as people may have initially hoped. These three areas are explored below.

3.4.1 Weakness 1 - The conduct of POP

The conduct of POP largely relates to adherence to the SARA procedure and, in particular, the issue of analysis and specificity when dealing with the problem at hand. In Scott and Kirby (2012), the need to both get and train the right staff (Chapter 9), but also for enhanced analytical support (Chapter 17), is highlighted at great length. The conduct of POP requires appropriate knowledge, skills and experience to be delivered effectively, but because these skills are not required for the traditional response policing model, there is currently a lack of these skills in police forces.

To chronologically bookend this point a lack of analytical skills was identified by Goldstein as early as 1990, (Goldstein, 1990), and was still seen as an issue in 2016 (Scott et al., 2016). The review of POP in England and Wales (Sidebottom, Bullock, et al., 2020) concluded that "recurrent weaknesses in the application of SARA...concerned the depth and quality of problem analysis.". Additionally they also found that "43% of survey respondents said they did not have access to information necessary to perform effective problem-solving". That the crux of POP lies in the understanding of the problem at hand, yet the police forces that want to implement POP do not have those skill sets available in sufficient quantities, it is hardly surprising that the conduct of POP can be sub-standard. However, it is encouraging to note that it would largely appear to be a resourcing issue rather than a systemic POP problem, as where analytical resourcing has been sufficient, largely as a result of collaborations with academia, POP successes have been strong.

If some analysis could be automated, or partially automated, then at least one bottle neck to further implementation would be widened. As will be shown later, modern NLP techniques combined with ML have the potential to allow the rapid exploitation of police free text information. If this information can be shown to have utility in the POP process, it is likely to contribute to lowering the analytical burden for a successful POP implementation.

3.4.2 Weakness 2 - The Delivery of POP

> Baldrick: But this is a sort of a war, isn't it, sir?
> Blackadder: That's right. You see, there was a tiny flaw in the plan.
> George: What was that, sir?
> Blackadder: It was bollocks.
>
> Blackadder Goes Forth: Goodbyee

Despite the best intentions of a plan and an analytical strategy, if they are not thought through, formulated and tested against the practicality of delivery, they are bound to fail even on seemingly mundane issues. This is shown in Blackadder's explanation to Baldrick about the precarious peace treaties built before the First World war: they looked good, they sounded good, but had anyone really stress tested the plan to ensure it would work? Were all parties committed to the plan, especially those with influence? Were the available resources made ready?

What is striking, to a former military planner, is that while the analysts seem to be well catered for within the POP community, planners are not. SARA is an analytical framework for POP. It is not a plan or an implementation framework. If it was an implementation framework, there would be processes for deciding how to judge and select responses, how and when to synchronise events, resource planning stages, questions regarding control measures (not comparison studies, but deconfliction in space and time), methods to test the plan and communication strategies. These are all valuable and necessary elements of a plan, but they are missing from the POP literature.

Hinkle et al. (2020) cites many implementation issues with third parties to the process, but if these have not been carefully managed during the planning process, and failure on their part understood, then the plan was never robust in the first place. That POP is not seen as mainstream policing is almost certain to hamper the process, and the necessity to deal with immediate problems now is hard to combat. Strong leadership, good plans and a cast iron belief that slower burning strategies will eventually pay off are required for implementation to be conducted effectively.

3.4.3 Weakness 3 - The requirements for evaluation

Evaluations are rarely sexy, and sometimes the resource or the impetus to conduct the evaluations vanishes as the project advances. Pet projects that aren't producing the results can fade away, while other project outcomes are so *obvious* that an official evaluation is not needed. Indeed, it is not in the culture of many organisations, especially the police, to systematically review their performance (Goldstein, 1990). Proper evaluation will make the process much more efficient, and analytical products and expertise will have lasting benefits if the real mechanism of change and benefit are realised. The jump in robust studies between the two Campbell reviews is encouraging, and the focus on evaluation in the POP awards in the UK and US will push it further along. However, it is worth noting at this stage that just under a third of submissions for a UK POP award[4] did not include any formal evaluation (Sidebottom, Bullock, et al., 2020). Again, in this area, as above, the necessity for skilled practitioners to do the work and leaders to allow them to do the work (or even mandate them) are keep elements for unlocking progress.

Three areas of weakness have been presented above: conduct, delivery and evaluation of POP. However, in each of those, there can be identified cross cutting themes that, if addressed, would lead to better POP outcomes. Two of these themes are access to information and analysis (Sidebottom, Bullock, et al., 2020). Access to information and analysis are also two areas were computational power and modern information systems can alleviate the workload. Information retrieval from document bases can be made much more efficient (if in doubt, consider google and other web search engines (Manning, Schütze, & Raghavan, 2008)) and the information within those documents can now either be automatically extracted or summarised (Kumar, 2011). That is to say that the underlying conditions that are creating these problem for POP can be at least partially addressed by leveraging technological improvements.

[4] Tilley Awards

3.5 POP Conclusions

It has been shown that POP is a successful practice for reducing crime and broader harms to the public. Additionally, it has been shown that this success comes down to understanding a problem in great detail, utilising all available information to form a specific response. Although the process involves the wider community, the practice is currently led and used almost exclusively by police officers using police data while relying on their scarce civilian analysts for support. POP is successful, but it is not a quick swap for traditional policing, as the resources POP requires are not readily found in police forces in sufficient quantity, and the culture of the organisation is not geared for its success.

In his article Tilley (2010) sets the agenda for the next round of improvements for POP, saying that "The research and development agenda for POP is now that of improving its efficiency and reliability in producing the intended outcomes." This is where the practical focus of this project lies. As shown above, analysts and access to information are key for a successful POP implementation, but they are often hampered by a lack of resources, either in the form of analytical support or resources to review disparate information. This piece of research attempts to leverage new but existing ML techniques to partially automate the extraction of information from police crime notes in the belief that, by doing so, the analytical burden for the scanning and analysis phases of the SARA cycle can be alleviated to some degree.

Chapter 4

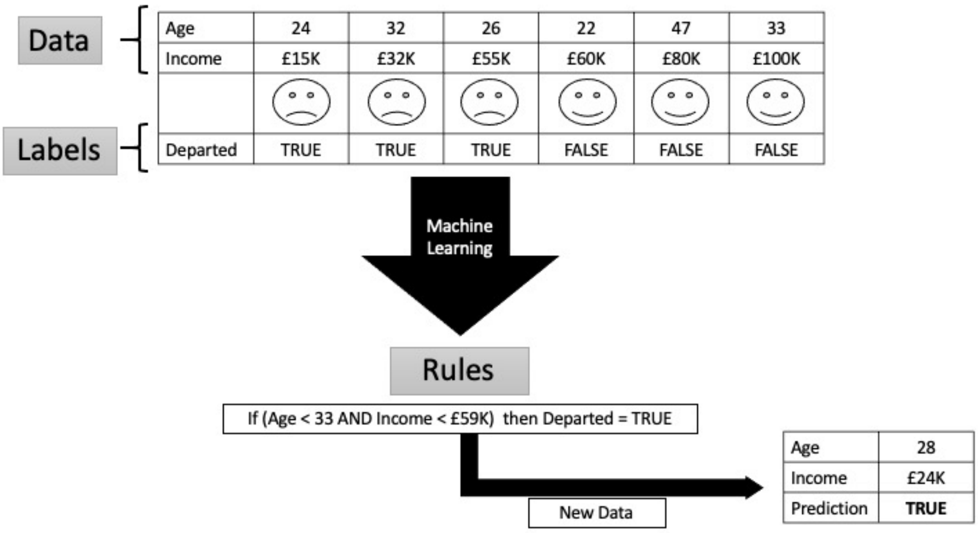
Machine Learning

“Machine Learning: Procedures for extracting algorithms, say for classification, prediction or clustering from complex data” (Spiegelhalter, 2019)

“With Machine Learning, humans input data as well as the answers expected from the data, and out come the rules” (Chollet & Allaire, 2018)

As shown in the quotes above, ML is about finding a set of rules or an algorithm that allows one to understand the structure of the data. That is, the overarching aim of ML is to discover a model or a set of algorithms or rules that assist in explaining the data. However, the real goal for those who use ML is often to take these rules and use the information that they provide against other sets of data to make predictions about unknown quantities. A toy example, Figure 4.1, is predicting customer churn in a business (Provost & Fawcett, 2013). Using historical data about customers, some known attributes like income and age can be used to try and explain the object of interest i.e. whether they have left the company. A ML algorithm can generate a set of rules to predict whether those historical customers left. These discovered rules can then, if the conditions are similar, be extrapolated to another set of data to predict who will leave in the future.

As shown above, the core of ML is about learning rules from data; however, the application of those rules to more data is generally where the main interest lies in the utility of ML. Once trained, these machines produce algorithms and rules that can then be used against unlabelled data to generate additional



Age	24	32	26	22	47	33
Income	£15K	£32K	£55K	£60K	£80K	£100K
	☹	☹	☹	☺	☺	☺
Departed	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE

Age	28
Income	£24K
Prediction	**TRUE**

Figure 4.1: A diagram to show machine learning. The inputs into the machine learning process are 1) the data and 2) the outcome of interest – commonly referred to as the label. The output from the process is a set of rules that can then be used to extrapolate to other, similar, data sets to make predictions. The rules can also be used to understand relationships within the original data. Adapted from Chollet and Allaire (2018) and Provost and Fawcett (2013)

labels at a reduced resource intensity but without, hopefully, a significant reduction in accuracy. In this way, ML generates rules, which can then be used to automatically extract information from wider sets of data. Extracting information with ML algorithms can be classified into two broad categories (though in truth, it is more of a continuum): 1) supervised and 2) unsupervised learning (Chollet & Allaire, 2018). Along the continuum between supervised and unsupervised learning is a process known as self-supervised learning, which is used to generate the models that this research leverages.

The next sections explores supervised, unsupervised and semi-supervised learning. With each type of ML highlighted with examples after each explanation. After these explorations into types of ML then data labelling will be explored, as labelling data is an important part of supervised learning. Finally limitations to ML will be explored particularly bias and explainability of ML models.

4.1 Supervised Learning

The key component of supervised learning is that the input data has already been labelled with information that puts the data into their desired class. Figure 4.1 is an example of supervised learning. For instance, a dataset of words may already have labels such as 'verb' or 'noun'. These labels are then used by the machine to being to build rules to classify the data inputs. The final ingredient required for success in ML is a measure of whether the rules are doing a good job or not. This can be a measure as simple as accuracy (what % were correct) to more complex calculations that can account for some permissible variation between given label and generated label. This success measure can then be used by the algorithm to select correct decision points and rules to improve the final rules. These final rules are then applied to unlabelled data, and the hope is that they are able to label the new data with similar accuracy (although almost certainly with lower accuracy). Chollet and Allaire (2018) identify four basic approaches to supervised ML, which are discussed below with examples:

Probabilistic Modelling. This style of model, of which Naive Bayes is the most widespread, attempts to find the probability of each potential classification *given the data inputs.* It is worth noting here that the input data for ML typically consists of a set of attributes (akin to explanatory variables) and, as previously mentioned, a data label (akin to a dependant variable). There are no real restrictions on the type of data that these attributes or labels can take. They can be discrete data, names or labels, or they can be continuous data such as numbers. However, the style of data one has will help to determine which algorithm to choose. Naive Bayes treats each attribute as equally important and independent from the other attributes, and using Bayes' theorem will calculate a probability for each potential label. The benefit with this algorithm is that the actual probability is not as important as the probabilities in relation to each other. That is, it is the relative size of the generated probabilities that is important, as the label with the largest probability is selected as the prediction. Another popular method in this class is logistic regression, which is also used to generate probabilities of a certain classification.

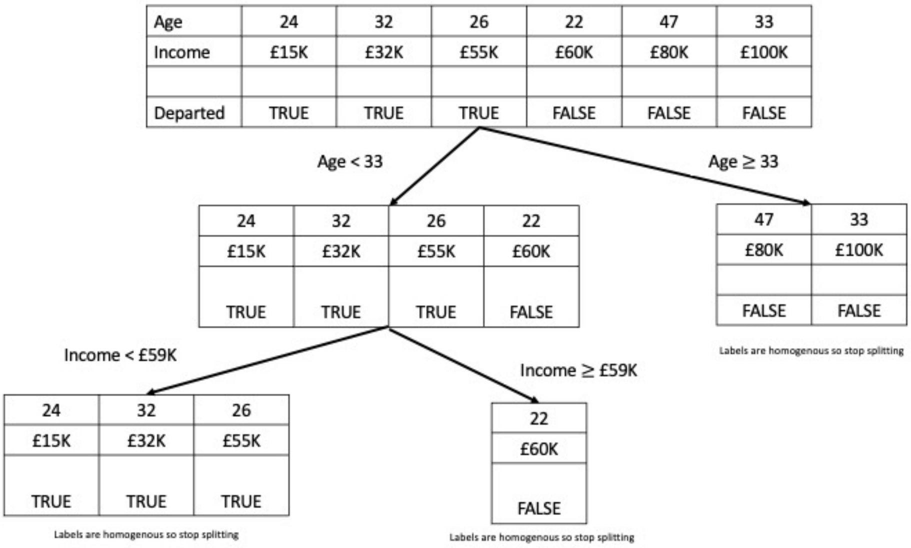

Age 24 32 26 22 47 33
Income £15K £32K £55K £60K £80K £100K
Departed TRUE TRUE TRUE FALSE FALSE FALSE
Age < 33
Age ≥ 33
24 32 26 22
£15K £32K £55K £60K
TRUE TRUE TRUE FALSE
47 33
£80K £100K
FALSE FALSE
Labels are homogenous so stop splitting
Income < £59K
Income ≥ £59K
24 32 26
£15K £32K £55K
TRUE TRUE TRUE
Labels are homogenous so stop splitting
22
£60K
FALSE
Labels are homogenous so stop splitting

Figure 4.2: An example of a simple decision tree. The tree first splits on the age of the customer, as that produces a homogenous grouping. The tree further splits the left-hand grouping by income, so that all groups are now homogenous. Source: Author generated

Divide and Conquer. This type of algorithm is typified by the decision tree. Here, an attribute is first selected, normally at random, and then the attribute is stratified to split the data with the aim of partitioning it as homogeneously as possible into sub-groups based on the given labels. These sub-groups are then further split by either the same attribute (but with different stratification) or by other attributes. See Figure 4.2 for a toy example. This can continue until certain conditions have been met, at which point the algorithm stops, and a set of rules has been generated. If the algorithm goes on for too long, there is a risk that each data point will have its own set of long and complicated rules generated. These rules may define the training data well but may not transfer well to other similar data that may need to be subsequently labelled. This process is known as over-fitting and is a flaw in ML algorithms that needs to be guarded against if the desire is to produce rules that are generalisable to new data. Overfitting is explored further in the limitations section.

In the case of decision trees, the maximum depth (i.e. number of splits or decisions) may be specified beforehand so that long and complicated rules can be avoided. This specification is an example of a hyper-parameter – that is, an additional guide to the formation of the algorithm that limits the possible set of rules that can be generated. Hyper-parameters can have a dramatic effect on the end result of an algorithm, and it is usually good practice to try a variety of hyperparameters when working through problems to explore result sensitivity. More sophisticated models in this class include random forests, which use lots of smaller, randomly generated decision tress and then combine them to produce a single result. The benefit of this is that it can avoid *local minima* whereby a normal decision tree is led by its procedure into a non-optimal path. More sophisticated still are gradient boosting trees, which only try to predict the actual data once, then spend the reminder of their time trying to minimise the residuals from the previous models with additional trees.

Kernel Methods. The classic example in this class is the support vector machine (SVM). This algorithm accepts the numeric data and maps the data to find a distinction between the groupings. For instance, if each piece of data consists only of two attributes, this can be graphed on a page (a 2-D vector space). Once all the data points have been plotted, a decision boundary can be formulated by finding a line that minimises the distance between itself and the two groups of data. See Figure 4.3 for a graphical example. Data with more than two attributes uses the same process but in higher dimensions of vector space.

The name kernel comes from a statistical process that reduces the computational power required by not requiring the plotting of all points in a vector space, but rather by allowing the distance between all point to be directly computed. This allows a swifter decision boundary formulation. SVMs can be susceptible to overfitting, and they have hyper-parameters that can balance the amount of misclassified instances with the simplicity of the computed boundary, which again leads to better generalisability from the model.

Neural Networks and Deep Learning. All the above models are considered shallow, meaning that they only carve the input space into very simple regions and find it difficult to pick up on underlying features in the data that should be invariant to simple changes. In the simpler shallow ML algorithms, this means that quite often, features have to be extracted or computed from the data, typically using expert domain knowledge. As an example, a list of profanities can be used to judge if a comment is suitable to be published or not.

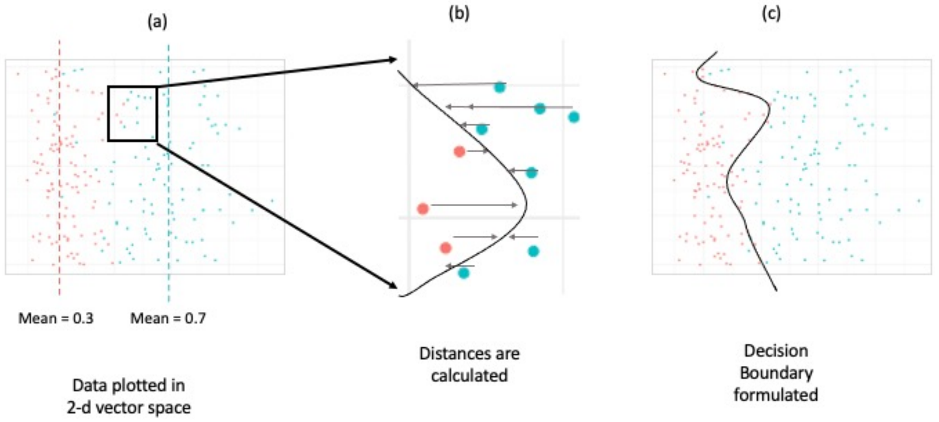


Figure 4.3: An example of a how a kernel method will split the data in a 2-dimensional example. Examples with higher dimensions are much more difficult to depict on paper but work in the same way. Source: Author generated.

With deep learning and neural networks, layers of models can be stacked that automatically form features in the learning process, passing these features from one layer to the next and therefore skipping the need for time consuming feature engineering. Deep learning has produced some remarkable results across a host of ML tasks in recent years and is seen as one of the most powerful ML tools (LeCun, Bengio, & Hinton, 2015). However, for this remarkable performance, a higher cost needs to be paid in 1) the availability of training data (typically neural networks require more training data then simpler models), 2) computational power (they can often need specialist hardware to produce timely results) and 3) model explainability (often the process of decision is hidden within the model and can be difficult to extract). However, as these models are further developed, some of these higher costs are inevitably lowered as they become the focus of more research.

4.2 Unsupervised

Unsupervised learning works in a different way than supervised learning. The algorithms attempt to ascertain the inherent structure of the data without any data labels. Unsupervised learning essentially attempts to group separate pieces of data according to the similarity between individual data points. The algorithms then split the data into similar groupings using these similarity measures so that similarities within groups can be identified. Two important methods of unsupervised learning are dimensionality reduction and clustering.

Dimensionality Reduction. Data can have many explanatory variables and attributes, and their values are unlikely to be independent of one another. Dimensionality reduction can combine variables using different weights to help condense the amount of variables (or dimensions) in the data to make it clearer what the most important ones are. Principal component analysis is a popular method for dimensionality reduction that has its roots in the mathematical community. Essentially, this technique recombines all the data in such a way that its dimensions are newly aligned to explain the most variation. Thus, by picking the most important new directions, the data set can be understood in a smaller number of dimensions or variables without significant loss of information. The trade-off is that not all the variation in the data is used, but what is used can be more easily explained and so the underlying causes understood.

Clustering. Perhaps the most popular unsupervised technique is clustering. Clustering seeks to group the data into different regions given its attributes. One of the most popular clustering algorithms is k-means clustering, which seeks to cluster the data into k different clusters. The algorithm works by selecting k random points in the vector space (the vector space dimensionality is defined by the number of attributes or explanatory variables), then computing distance measures to allocate each data point to a group. Group centres are then recalculated, and distances remeasured, and this continues until the tightest clusters are discovered. The k, how many clusters to use, must be provided to the algorithm at the outset, but is typically not known. k can either be found through running variants of k and finding the 'best' one or by using hierarchal clustering or expert knowledge. Once clusters have been found, these are then explored to deduce statistical characteristics, or as mentioned above, they can be combined with other data to provide a richer picture.

4.3 Semi-supervised Learning.

In between supervised and unsupervised learning is semi-supervised learning. This is a type of learning that uses labelled data, but the data has not been labelled by humans. Typically, the label is known because it is inherent to the data. For example, semi-supervised learning on text data can occur through word prediction. A complete sentence has a word randomly chosen and masked, the machine is then given the sentence, complete with the word gap. and it must guess the masked word. The masked word has a label – its actual value – and so it is supervised learning, but the label has not been generated by a human, so it is a much less laborious process. Later, it is shown that semi-supervised learning is one of the pillars that has led to the production of PTMs by allowing models to efficiently learn from huge datasets with little human intervention.

4.4 The Labelling Burden.

The key difference between the two main learning methods outlined above is data labelling. Labelling data is not a trivial endeavour though it is often worthwhile. Castelli and Cover (1995) have shown that labelled data examples are worth exponentially more than unlabelled examples (that is, in certain circumstances, they are able to reduce the probability of error exponentially over the same number of unlabelled examples), so even though they are more difficult to come by (they will almost certainly require resources to generate), it is often worth labelling data to achieve a better outcome in the long run.

However, labelling requires an initial investment of resources, investment in a model that may not work or produce the results wanted. Also problematic is labelling data for fluid problems. What may seem like valid data labelling initially may no longer be so after the problem has morphed. This problem is not new, and many scholars and practitioners have been at work trying to lower the labelling burden. As can be seen in Figure 4.4 there are a number of strategies that can be employed to reduce the labelling burden, with trade resource utilised for overall accuracy. These methods are explained below.

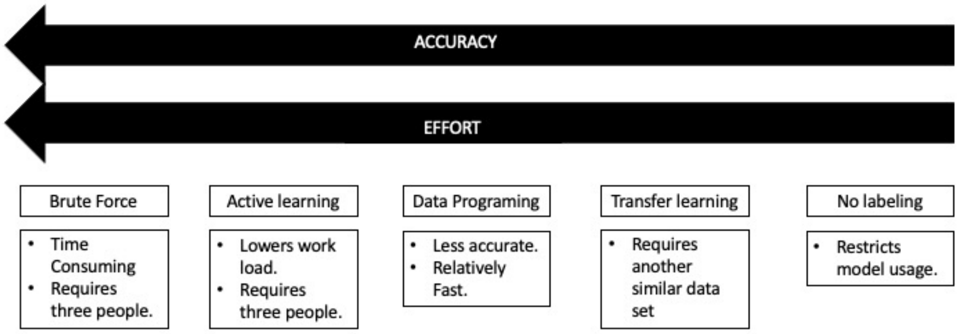


Figure 4.4: A Summary of different labelling strategies. Source: Author generated.

Brute Force. This is hand-labelling all the data required. This will include the training set and the test set. It is normally done by humans, who can be employed in a variety of ways. Depending on the subject matter expertise required. The cost of labelling can vary considerably. The skills to label x-rays of fusions in spinal surgeries are almost certainly rarer, and therefore more expensive, than the ability to decide if a tweet is offensive or not. Humans are also not infallible, and they can be subject to biases (Kahneman, 2011), meaning that generally enough people need to be involved to gain a consensus – typically, this means at least three, but some datasets have employed more. However, the brute force system is generally the most accurate of all the measures[1] - a fine luxury if you have the resources.

Active Learning. "The key idea behind active learning is that a ML algorithm can perform better with less training if it is allowed to choose the data from which it learns." (Settles, 2009). So how does a machine choose which data to learn from? Essentially, the machine is fed a small amount of labelled data, far less than one would hope to use in the normal run of things. The machine learns from this seed data and then assigns a probability to each unlabelled data point, and a decision boundary is formed. Those data

[1] This relates to "out of the box" functionality, some data sets have been more accurately labelled by trained ML algorithms, (see https://rajpurkar.github.io/SQuAD-explorer/), but of course the models were first trained on human labelled data.

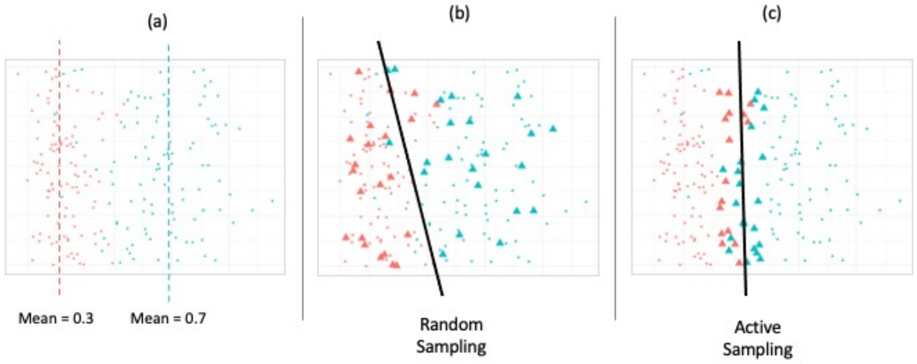


Figure 4.5: Panel (a) shows two 1D normal distributions with means 0.3 and 0.7. Panel (b) is the same distributions highlighting those labelled with a random sampling strategy, and the thick black line is a plausible decision boundary. Panel (c) is the same distributions, but now the labelling has been completed in accordance with an active learning strategy. The thick black line is a plausible decision boundary based on this method. Source: Author generated

points that were difficult to decide upon, those that were close to the decision boundary, are then chosen for labelling by a human, and the cycle is repeated. See Figure 4.5 for a simple example. The benefit, as can be seen in Figure 4.5, is that each actively labelled data point contributes much more information to the formation of the decision boundary than those selected at random. Selecting points far away from the boundary generally has little effect on the decision boundary, and so for the same labelling resource, less information is achieved. While this is a simple one-dimensional example, it can be scaled to more complex environments with more sophisticated techniques, but the principles remain largely the same.

Transfer Learning. "Transfer learning is used to improve a learner from one domain by transferring information from a related domain." (Weiss, Khoshgoftaar, & Wang, 2016). Transfer learning is centred around using the knowledge gained from one data set, usually in the form or algorithmic rules, on a second, related data set. Typically, there is a resource hurdle for labelling the second data set that can be lowered by utilising the information from a data set that has already been labelled or curated such that the accuracy is known to be high. Examples of this include utilising language algorithms generated for one police force to help label the training data to be used with a second police force, or as we will come to see, transfer learning can also play an important part in key NLP model steps such as PoS tagging and word embedding, where a word is represented by a vector of numbers that reflects its similarity to other words.

Data Programming. Data programming is a form of weak supervision where knowledge is used to guide the labelling of data through the application of heuristics or simple rules. Snorkel, (Ratner et al., 2017), is an example of this type of modelling that takes simple rules developed by SMEs, then combines and weights these rules to automatically produce labels for data points. An example of a simple rule might be *Text contains "victim knew offender"* or drawing on a dictionary of known relationships (dict:relationships) the rule might be *Text contains "Offender is victim's (word in dict:relationship)"*. These rules are not tested against labels, but each other to identify where there is agreement and correlation (too much correlation is bad as it essentially over emphasises the same relationship), rules are then weighted and labels generated. It was found in Ratner et al. (2017) that time spent generating rules was much more efficient than time spent labelling data, but that did depend on subject matter expertise and rule writing proficiency of the individual rule authors.

In summary, labelled data for ML algorithms is a good thing and can be exponentially beneficial for providing the information sort. However, it is difficult to come by, especially in niche fields where the skills needed to label the data are scarce. Other fields where the questions are more fluid will also encounter labelling issues as, potentially, the data set has to be re-labelled for each purpose, unless the underlying representations can be unlocked. However, a body of research that is developing techniques to lower the labelling burden, without much reduction in overall accuracy, is encouraging. There will always be a requirement to label some data – if only to test that the model is working correctly – but speeding up the process and lowering the hurdle for entry will enable more powerful ML techniques to be used.

4.5 Predicting Performance

Once a ML model has been trained, it is generally then tested on unseen data to understand how good it will be on unseen instances of data. For this research, the models will be used for classification tasks, and so prediction performance will be explored here in that context.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4.1}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{4.2}$$

$$Precision = \frac{(TP)}{(TP + FP)} \tag{4.3}$$

$$F1 = \frac{(2 * TP)}{(2 * TP + FP + FN)} \tag{4.4}$$

$$MCC = \frac{(TP * TN – FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.5}$$

Where: TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

The simplest form of metric for predictive performance is Accuracy (a capital “A” is used to differentiate the metric from the everyday usage). The equation for Accuracy is given in Equation 4.1. Essentially, it is the percentage of all correct predictions divided by the total number of elements to be predicted. Accuracy is easy to understand but can sometimes conceal poor performance when the dataset is imbalanced. An imbalanced data set is where one of the classes to be predicted is rare in relation to the other class. For example, imagine trying to predict a crime like domestic abuse when only 1 in 100 crimes are domestic abuse. A classifier that pays no attention to the data and classifies everything as not-domestic abuse would get an Accuracy of 99%, which is high, but the model is poor, because it will never find any domestic abuse crimes.

When the data has imbalanced classes then it is important to use other metrics in place of Accuracy. Researchers have found two metrics that are useful to tracking classification tasks, “Precision” and “Recall” (Witten, Frank, Hall, & Pal, 2017, Chapter 5). Precision (Equation 4.3) is a measure of the relevant instances amongst all the retrieved instances and Recall (Equation 4.2) is a measure of how many relevant instances were retrieved. These two measures typically tend to be inversely related as selecting more of the relevant instances increases the chances of selecting irrelevant instances. For that reason the F1 measure was developed (Equation 4.4), this takes the harmonic mean of the recall and the precision and is therefore a combined measure of both of those metrics.

Further research (Chicco & Jurman, 2020) has shown however that the F1 measure can still be misleading and that a more intricate measure - The Mathews Correlation Coefficient - can be more effective at discriminating between classifiers. The major differences between MCC and the F1 score is that MCC is invariant to class change (so if the classes are swapped there is no change in the MCC metric) and secondly but relatedly the F1 score does not account for True Negatives (classifying irrelevant instances as irrelevant). For these reasons the MCC metric will be adopted throughout this research as the primary means of assessing model performance.

4.6 A General Approach to ML

Having introduced the various aspects of ML this section will specify the general approach for supervised ML as that will be the approach used throughout this research. The approach is therefore as follows:

1. Split the data. The data is randomly split into three sets: train, validation and test. The train set is the data that the model will be trained on. The validation set is used to help select the correct hyper parameters for the model. The test set is the data that the model performance is judged upon after the final model selection.

2. Label the data. All data in each set are read and labelled by human annotators.

3. Train the model. The model is trained on the labelled training data. Hyper parameters are selected, and the effects are judged using the validation set. In a sense, the validation set is an intermediary test set that helps select hyper parameters.

4. Test the model. Once the model has been trained with the final hyper parameter selection, the test set is predicted, and the model-generated labels are compared against the human labels to judge the model performance.

4.7 ML Limitations

ML has seen a surge in utility in the last decade or so as processing power and data sets have become increasingly available. However, it is not without some drawbacks and issues that can hamper its effectiveness or utility in certain scenarios. Some of these major limitations are explored below.

4.7.1 Overfitting.

“The fundamental issue in ML is the tension between optimization and generalisation” (Chollet & Allaire, 2018). Overfitting in ML is where the algorithms have been optimised for the training data, but in doing so have over generalised their rules to the variation in the training data. By doing so the ML model has therefore lost some of the prediction power on the data set in general. Every data set has some natural variation, this natural variation is variation in the data that is derived from explanatory variables that are not in the model or interactions of existing variables that are not modelled correctly. When a model over fits, it is essentially predicting this variation from the existing model, but without the mechanisms or information to do so, so it is learning incorrect relationships.

Figure 4.6 shows pictorially how this may occur, the distributions in panel (a) are random samples from two different normal distributions with separate means and standard deviations. Predictably, there is an overlap between the points, but knowing the distributions makes it possible to mathematically deduce, using probability theory, a decision boundary that will map a line

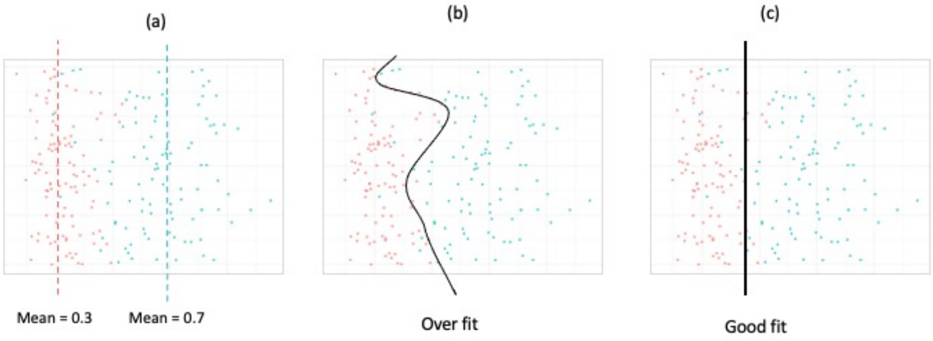


Figure 4.6: Pictorial example of Overfitting. Panel (a) shows two 1D normal distributions with means 0.3 and 0.7. Panel (b) is the same distributions with an overfitted decision boundary (black line). Panel (c) is the same distributions, but the thick black line is a plausible decision boundary based on the known distributions (it is slightly left of 0.5 as the red class has a lower variance). Source: Author generated.

whereby on one side of the line, the probability of a red data point is higher than that of a blue, and on the other side, the converse is true. That is, the optimal decision boundary is known. However, in general, the ML algorithms do not have the specified distributions and have to fit on the data provided. Therefore, depending on how much the algorithms value getting every data point classified correctly over the simplicity or generalisability of the rules will depend on how susceptible it is to overfitting. Some techniques to prevent overfitting include the following:

1. Have a test set. It is best practice to split available data right at the outset into a test set and a train set. The train set is set aside and is only used at the end to evaluate performance on the chosen model. It is not used to train models or select models.

2. Get more data. The more data one has, the more likely the true patterns are to be found.

3. Divide the data. A typical technique here is cross-validation, whereby the train data is randomly split into, typically, ten different groups, and

then the model is trained on nine of these groups at a time (a different group of data is left out on each occasion). The resulting models are then tested on the left-out data group, and the results compared and analysed to pick the best generalising model.

4. Restrict the model. Do not allow the model to form overly complex rules. This can take the form of only allowing so many branches on a decision tree or by requiring a certain smoothness to a decision boundary in a probabilistic model.

4.7.2 Explainability.

"In general, humans are reticent to adopt techniques that are not directly interpretable, tractable and trustworthy." (Arrieta et al., 2020). Being able to understand how an algorithm works is important for several reasons, primarily among them being the trust that the end user will place in its predictions. The ability to understand why a decision is made greatly increase the confidence in it. Understanding how a decision was made can also have additional benefits, including ensuring impartiality of decision making, robustness to new data and identification of causality between the variables and the resulting class.

Explainability is relative to the audience: what might make sense to one person may not make sense to another. In general, if a problem is complex, then more complex algorithms lead to more accurate predictions (Arrieta et al., 2020). This has obvious implications for those who wish to use ML, who have complex problems but also a mandate to understand how the predictions were formulated and what bias, if any, are in the system. A field called explainable artificial intelligence (XAI) (Gunning et al., 2019) has developed to try and quantify these questions and develop a suite of tools to aid the model builders and the users in understanding their predictions better. However, as the authors of Gunning et al. (2019) acknowledge, how to reliably and consistently measure a good explanation is still an open research question, not least because the standard and style of the explanation can differ between intended audiences for the same model as well as across models and domains.

In his seminal paper on explanation in AI (Miller, 2019) Miller gives four major factors for good explanations. First, explanations should be contrastive – they should explain the output of a single instance by contrasting with hypothetical counterfactual cases. For text, this could be changing words within the sentence. Second, the explanation should be selective: the explanation should not try to list every cause of a generated output, just the most important. Third and perhaps the most upsetting to a statistician, "probabilities probably don't matter" , referring to probabilities is less impactful than referring to causes. Lastly, Miller states that explanations are social, and thus they are contextual relative to the understanding and competence of the explainee.

A popular tool for interrogating ML models is LIME, (Ribeiro, Singh, & Guestrin, 2016). LIME builds a simpler local model around a prediction to help draw out the locally important factors for a single instance of data classification. These individual models can then be aggregated to provide a view across a larger dataset. LIME will be used in the studies within this research and is explained more fully in the Methods chapter. This tool relies on the contrastive model as set out by Miller (Miller, 2019). The output of the model can be adjusted or presented in different ways so that the remaining elements of a good explanation can be met, in particular tailoring the explanation to the audience.

4.7.3 Bias

ML systems can have bias making them unfair, where unfairness is "prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics." (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021).). Clearly, bias in a ML system is sub-optimal, as it can lead to groups being discriminated against and a reduction in trust in that model and other AI systems. Bias in ML systems stems from two main areas: the data and the algorithm. These two main areas are explored to see how they may introduce bias.

Data Bias

Data bias can come from a number of different sources, the most important being the following two.

1. Representation bias. This can be where the sampling of the population has not been completed in a representative way. Police data suffers from this bias, as recorded crime is not recorded uniformly across victim and crime types (**baumer2002neighborhood** ; **tarling2010reporting**).

2. Omitted variable bias. This occurs when important variables are omitted from the data. Within police text data, this could be observed if certain events are not mentioned in the texts to be analysed.

Other sources include aggregation bias, where rare but distinct groups have inferences drawn about them that are derived from population characteristics, and measurement bias, where the quantity and quality of measurement can vary between groups.

Algorithmic Bias

“Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm ” (Mehrabi et al., 2021). That is the bias is introduced by the choices the researchers makes in the selection of model types and parameters (Hooker, 2021).Some models are better at some problems than others. Tuning hyper parameters are also likely to bias in favour of correctly predicting certain instances over others (Paiva, Moreno, Smith-Miles, Valeriano, & Lorena, 2022). In relation to crime text data, there may be rare words in certain crime types that may not be represented well with the models chosen and therefore may lead to inappropriate classifications. This could introduce bias with certain crime or victim types by the selection of the algorithm.

Measuring bias

We have seen that bias can stem from two main areas, the data or the algorithm. Similarly bias can present in two main areas (Chouldechova, 2017). Firstly predictive accuracy, do the results from the ML system have the same accuracy across different groups? Secondly when the ML system makes mistakes are those errors equally likely across different groups of people.

A well studied example of bias in the literature is that of COMPAS a system used in the USA to predict recidivism rates. This example also allows a better understanding of the two different biases (Chouldechova, 2017; Kleinberg, Mullainathan, & Raghavan, 2016). ProPublica, an investigative journalism group, produced research that showed that error rates with the COMPAS system meant that members of the black population were more likely to be misclassified as high-risk offenders, and white people were more likely to be misclassified as low-risk offenders (Jeff Larson, 2016). Northpointe, the providers of the tool, countered this claim with evidence that showed that the accuracy for prediction across racial groups was similar, in that regardless of racial group, the accuracy of predicting high or low recidivism rates was the same.

Further research (Chouldechova, 2017; Kleinberg et al., 2016) not only showed that both pieces of evidence were true, but that they were almost inevitable in a system where the underlying rates are different between different groups (in this case the data used (itself not without inherent biases), has different recidivism rates for the white and black populations). Therefore, in one sense, there was no bias, because the COMPAS system had the same accuracy across racial groups. However, when looking at the second source of bias, the errors, it was shown that the system was biased, as the direction of the errors was different for the two racial groups, with black people being more likely to be classified as high-risk offenders when they were not and therefore subject to more punitive measures. However, as shown in Kleinberg et al. (2016), with underlying differences in the recidivism rate for the two groups an unbiased error rate is not possible (except in the case of a perfectly accurate system).

So what? First, measuring bias is not straightforward and looking at single measures can skew interpretation. Second, understanding the impact of the bias is also crucial, as inaccurate predictions in one direction can be more costly than in another direction. Third, where different underlying rates are recorded a perfectly unbiased system is not possible in practice (Kleinberg et al., 2016). An excellent overview of this problem and its interpretation is given in (Hellman, 2020). For this research and measuring the bias, I will therefore measure both the bias in accuracy and the bias in error rates. The two metrics are *predictive parity* and *equality of outcome* respectively. These metrics will be formally introduced in the methods chapter.

4.8 Summary

This chapter has introduced the broad concepts surrounding ML. The chapter explored the main paradigms of ML, how they operate, what they require and how success is measured. Important limitations for ML include overfitting to the training data, the degree to which models can be explained and any biases they may contain. The research in this thesis is largely based on supervised learning and so requires labelled data. Active learning is used to label the data, and performance is judged through the MCC metric. Further applicability of the models is explored by using the LIME tool to explain how the models came to their decisions, and bias metrics are used to explore bias in the system.

The next chapter moves on to a specific section of machine leaning, NLP. NLP is used when the data to be analysed is textual data. The next chapter takes the concepts explored here and shows how they can be built upon for analysing free text data.

Chapter 5

Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence that seeks to extract information from text, principally free text. The main purpose of NLP is to "make human language accessible to computers" (Eisenstein, 2018). This is generally accomplished by accepting data as words and then representing them in the form of numbers. Once the data is in the form of numbers it can then be manipulated by the ML processes outlined in Chapter 4.

Within NLP there is a tension between knowledge (trying to understand the structure of language) and learning (using algorithms to efficiently code representations with a focus only on the results) (Eisenstein, 2018). The knowledge advocates have been working on language problems for some time under the guise of computational linguistics. They have laid down many important NLP foundations that can be used to automatically extract information from text, such as part of speech tagging and parsing a sentence so that the dependencies between words can be understood. As Manning (2015) notes, there has been a shift in NLP more recently to those with more of a focus on the learning. That is, they want to use modern machine learning processes, and especially deep learning processes, to translate raw text directly into the desired output (Eisenstein, 2018).

This research aims to leverage this shift in research focus and utilise the most effective NLP models based on the learning construct. In particular to utilise a class of models known as PTMs. PTMs, introduced earlier, are powerful

NLP models because they already have an element of language understanding built-in before they are used on a specific task.

The analogy used in the introduction was PTMs were like employing a graduate and conducting job specific training, rather than earlier NLP models which required full training in an area before they could be used. Most of these PTMs have been pre-trained on edited text, such as news reports or other structurally published material. For that reason, the models have been built on text that is likely to have a different compositional structure than police free text data, and so PTM utility with this data is not obvious.

Figure 5.1 gives a generic language processing pipeline that might be used to gain information from free text. The focus of the previous chapter was on the machine learning models to the right of the diagram. This chapter, however, comes before machine learning in the process and is concerned with the processing of the data – that is, finding an appropriate representation of the text in the form of numbers.

A few notes on terminology. A token is an individual element of interest, which generally will be a word, but it can be a piece of punctuation. It is the lowest level of investigation. Tokens are collected to produce a document. A document in this research is a single description of a MO for one crime; however, other examples are a single tweet or, in certain cases, a whole book. A collection of documents is called a corpus. In this research, the corpus will be a collection of MO data, all from the same police force and of the same crime type.

This section begins with a note on different applications for NLP, then moves onto techniques that can be used to harmonise and understand the individual tokens in a document. The section then progresses to how these tokens can be represented within a document and how the differences between documents within the same corpus can be mapped. Finally pre-trained language models (PTMs) will be introduced. PTMs are the most advanced kind of NLP model and are the model that this research will be based upon.

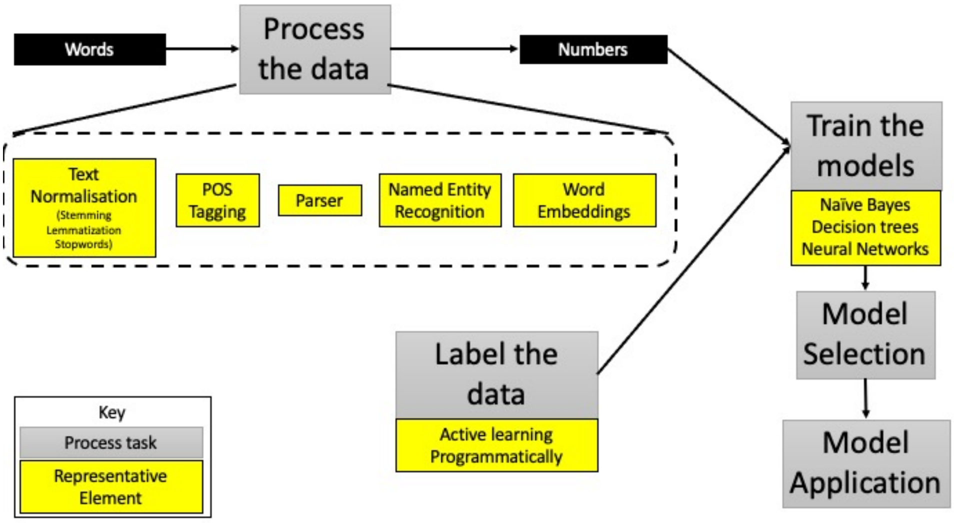
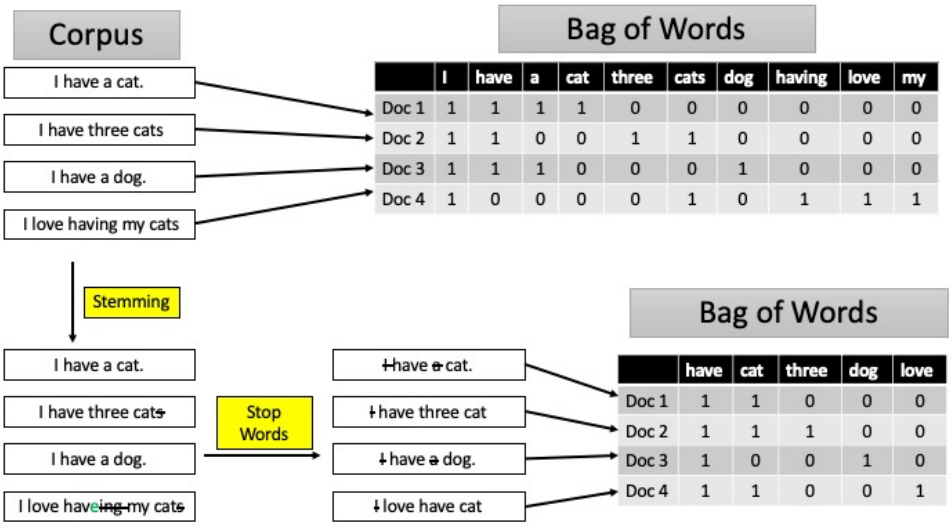


Figure 5.1: This figure demonstrates a generic machine learning task containing NLP. The exact details of the tasks will be explored in this chapter. Source: Author generated

5.1 Applications

Natural language processing has many applications, just like machine learning in general. However some important applications are as follows:

- Classification. Classifying documents into one of several categories can be general, such as a positive movie review, or more specific, such as a MO where the offender has used a knife.
- Information Extraction. This may be to extract the disease from clinical notes, without knowing exactly what disease it is you are searching for.
- Question and Answering. In this application questions are asked of a specific corpus and the answer is returned. In this application both the question and the corpus may need to be subjected to NLP techniques to generate the answer.
- Translation. Translating form one language to another.
- Chatbots. Where computers are designed to respond to conversations with humans.



	I	have	a	cat	three	cats	dog	having	love	my
Doc 1	1	1	1	1	0	0	0	0	0	0
Doc 2	1	1	0	0	1	1	0	0	0	0
Doc 3	1	1	1	0	0	0	1	0	0	0
Doc 4	1	0	0	0	0	1	0	1	1	1

	have	cat	three	dog	love
Doc 1	1	1	0	0	0
Doc 2	1	1	1	0	0
Doc 3	1	0	0	1	0
Doc 4	1	1	0	0	1

Figure 5.2: This figure demonstrates how a bag-of-words algorithm operates. The tokens in each document are counted and a matrix is formed with a column for each word in the corpus and a row for each document in the corpus. If document 1 contains the word in col 2, then a 1 is placed in the cell (2,1). Two techniques to provide a mores succinct output are also shown. Stemming, which reduces word forms, and stop word removal which removes common words of little value. Source: Author generated

For this research the focus will be centred around classification, as it is generally considered a gateway task before moving onto more complex applications.

5.2 Text Normalisation

On of the most simplest forms of NLP output is what is known as the bag-of-words modelling. This method produces, an unordered, representation of all the words in a given document by producing a matrix with *1* if the word is present and *0* if the word is not present, Figure 5.2 gives a toy example. This matrix, called a word-document matrix, can then be used as input to machine learning algorithms. Some elements to note here are that the order of words is not kept, so some of the semantics of the language can be lost. Secondly, with slight different variation in word forms, even a small selection of basic sentences can give rise to a relatively large matrix, this makes it harder for the computer to grasp the meaning of the words, is *cat* really that much different from *cats* that they need separate columns? Reducing the size of the matrix will also make the computation more efficient as there is less matrix manipulation to conduct. What follows is a brief exploration of the techniques to reduce the variation in tokens within a document to help convey the same or very similar meaning with less tokens.

5.2.1 Stemming

Stemming is the process of removing inflectional affixes from a word. Examples of inflectional affixes include the plural marker *s* and the past tense marker *ed* (Eisenstein, 2018). See Figure 5.2 for an example. Stemming groups words with the same underlying concept and so reduces the total number of different tokens in a document and corpus, allowing similarities between tokens to be more easily identified. As an example horse, horses and horsed all become hors once stemmed. Note the stem in this instance is not a real word, but this would not affect the algorithm (Jivani et al., 2011). Stemming is conducted through a rules based system, and so on some occasions, the stemming generated is not correct. An error can either be over-stemmed, where two words of differing meaning are given the same stem – for example, Williams to William – or it can be under-stemmed, where two words with the same meaning have different stems – for example, tooth and teeth.

5.2.2 Lemmatizing

Lemmatization is an additional step that can be used to reduce the amount of individual tokens in a word-document matrix. It is similar to stemming but attempts to avoid some of the pitfalls of the rule-based system by additionally understanding the context of the word. One of the more popular lemmatizers, WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), is a lexical reference system, similar to a database or dictionary, which lists words and there synonyms under a joint lemma, this means that WordNet can be interrogated with a given word and its part of speech, such as noun or verb, and after a look-up, a lemma will be returned. The system can make less mistakes than stemming, as the rules are hardcoded; however, computationally it can be more expensive.

5.2.3 Stop words

Stop words are words that are so common (e.g. *the*, *a* and *to*) that they are generally thought to play little role in the linguistic meaning of a document. Stop words are corpus dependant, so while there are lists of common stop words it is best practice to tailor each list to the text at hand. As an example we can see that even after the removal of some classic stop words in Figure 5.2, *have* appears in all documents. If this was a real example then there would be serious consideration for including *have* as a stop word as it does not assist in the discrimination between documents.

Normalising text can have its advantages. The meaning of a document can be distilled to a smaller size, and additional rules and dictionaries can be leveraged to clear some ambiguities. However, removal or changing of a token can reduce the information that is left in a document, information which may be useful for further discrimination of the NLP model. Bag-of-words models in general lose all their semantic value as the word order is lost, and so these techniques are particularly useful for those situations where semantic importance is not high.

5.3 Word Features

5.3.1 Part of Speech Tagging

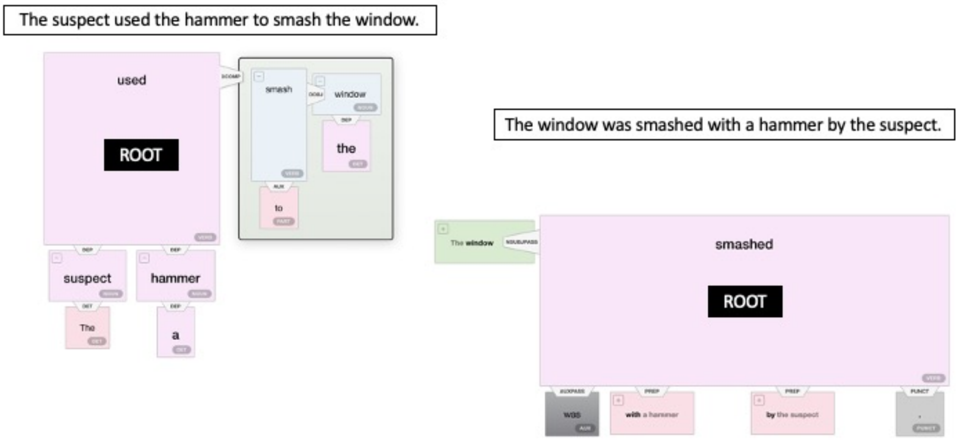
Another method for enriching the data set is to understand what part of the syntax of a document each token represents. That is, a part of speech (POS) tagger will label each word as a noun, verb, etc. By labelling the words in this way, some ambiguous meaning can be avoided. Take the following headline as an example: "Dealers will hear car talk at noon". If *talk* in this example is a noun, then there are no surprises; however, if it is a verb, then we may question what kind of dealers they are and what they have been doing with their produce. A popular and effective (Zeman et al., 2018) open-source tagger - UDPipe (Straka, 2018) – can label an English sentence with the correct POS tags with around 90% accuracy. These models have been built using labelled data from the universal dependencies treebank. In this tree bank 37,000 English sentences have been hand-labelled with their POS tags. This labelled data has then been used to generate the open-source model UDPipe. Part of speech tagging is useful for understanding text in general; however, it can have more specific uses, such as finding entities like names or addresses, in a process known as named entity recognition (NER). This is described next.

5.3.2 Named Entity Recognition

As the name suggests, NER helps to draw out information from text relating to real world entities, be they people, places or organisations (Eisenstein, 2018).More recently, the types of subject entities have been widened to include drugs, medical conditions and different types of biomedical items such as protein types (Goyal, Gupta, & Kumar, 2018). NER is an important step in many NLP applications because it helps to draw out salient information between document types. These named entities, once discovered, can form part of the feature engineering of a document and be linked across documents as additional information. A popular open NER model is the Stanford NER (Finkel, Grenager, & Manning, 2005). This model was trained on newspaper reports that have been manually tagged. The data a model was trained on will have implications for its use outside of that domain. As Prokofyev, Demartini, and Cudré-Mauroux (2014) show models trained outside of highly specialised domains show significant drops in their effectiveness, as such testing of open-source models and possibly adaption is required on new styles of corpus.

5.3.3 Sentence Parsing

An additional measure that can be taken with sentences within documents is to parse them so that the internal dependencies are understood. Dependency parsing takes a sentence then produces a dependency for that sentence, beginning at the root of the sentence and cascading to all words within it. The root is decided by a set of deterministic rules dependant on the type of sentence and the word types within it (Eisenstein, 2018). Two examples of a dependancy tree, the result of sentence parsing, are shown in Figure 5.3. Knowing the dependencies between words is useful, both for information extraction but also question and answering tasks, because understanding the underlying dependencies in a sentence can help clear some of the ambiguities that were introduced from the manner in which it was presented. The clarity sentence parsing can bring is seen in Figure 5.3 where the two sentences,



Dependencies from https://demo.allennlp.org/dependency-parsing

Figure 5.3: This figure demonstrates the dependancy parsing of two similar sentences. Both parses produce a similar structure of dependancies despite the difference in the wording. *suspect* and *hammer* both join to *smash(ed)* before they reach the *window* in both parses. Source: Author generated

with essentially the same meaning, have very similar dependancies despite the difference in wording. We see that both the *suspect* and the *hammer* have to pass through *smash(ed)* before they reach the *window* in both cases. An additional important application of sentence parsing is negation, whereby it is important to tract where a negative clause is acting.

5.4 Word Representation Methods

> You shall know a word by the company it keeps
>
> J. R. Firth

In this section so far there has been an introduction to how to normalise the text by reducing the amount of individual tokens in a document, then building on that by understanding what features can be extracted by dependency parsing, NER and POS tagging. Explored here is the meaning of the individual words, how the meaning contributes to the totality of the document's meaning and how similar words can have similar meanings across documents.

Token	**TF**	**DF**	**IDF**	**TF-IDF**
pet	1	10	1	1
dog	2	50	0.3	6.65
cat	2	100	0	0
Fido	1	1	2	0.5

Table 5.1: Example TF-IDF values for a 100 document corpus.

5.4.1 Frequency Methods

The bag-of-words model utilises the simplest representation of words, the words presence or absence is noted by a binary marker (generally 1 or 0) see Figure 5.2. This method does not draw any explicit meaning from the word itself, it only marks its presence. The next stage up from this is to change the binary marker to a count maker, so that the number of times the token is present in a document is now recorded, giving additional weight to multiple uses, although as there is a tendency of words to cluster this method tends not to show much improvement on the simple binary choice (Eisenstein, 2018), this is known as term frequency.

An additional method utilising the same approach seeks to understand how important a word is in that document, given its prevalence in the corpus as a whole. This method is known as TF-IDF, which stands for term frequency – inverse document frequency. The first part, term frequency, was outlined above and is a count of the terms in that document. The second part, the inverse document frequency, is a measure of the number of documents that the word is mentioned in. It is inverted because words that are rare in the corpus should have more discriminatory power (Manning et al., 2008). Typically a logarithmic measure is used.

Each of the procedures outline above was demonstrated with single tokens; however these procedures can be generalised to groups of tokens which represent phrases. Groups of tokens are known as *n-gramms* where *n* relates to the number of individual tokens in a phrase. An example of a tri-gram is *New York City.* n-grams can either be used exclusively or alongside single tokens so that common phrases can be extracted for greater fidelity, especially useful, if as with the example above, the n-gram is a single entity.

The examples above have distilled the information from each token into a

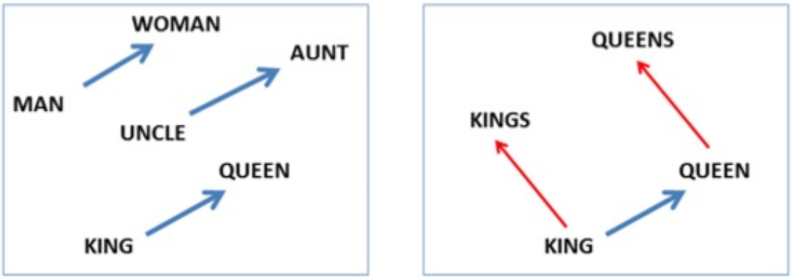


Figure 5.4: Left panel shows vector offsets for three word pairs illustrating the gender relation. Right panel shows a different projection, and the singular/plural relation for two words. In high-dimensional space, multiple relations can be embedded for a single word. Source: Mikolov, Yih, and Zweig (2013)

document down to a single number, and the presence or size of that number contributes to the meaning of that document within that corpus, along with the distribution of the other tokens. That single number represents that word. What this process still not allow for, however, is the individual meanings of words, as opposed to their mere presence, to contribute to the characterisation of the document. For this reason, word embeddings were developed that would more accurately contribute to the meaning of individual words. These are discussed next.

5.4.2 Word Embeddings

The general idea behind word embeddings is to represent each word with a vector of numbers (typically, they can go as high as 300 numbers for one word) such that words with similar meanings have similar vectors. Either these embeddings can be generated for individual corpuses, or previously derived embeddings can be used. Typically, these derived embeddings have been trained on a massive corpus, such as the whole of Wikipedia.

Embeddings are generated in a number of different ways, but in essence they exploit the same relationship given in the quote at the start of this section – that is, a word is defined by those around it. A single word of interest may occur in a corpus a number of times, but if it is conveying the same meaning, the words surrounding it will be similar. Embeddings are created by investigating the probability of seeing a neighbouring word, given the target word (Mikolov, Chen, Corrado, & Dean, 2013). Those target words with similar property distributions for the same neighbouring words, will have similar meaning. The property distributions are encoded into the vectors of interest and relationships and similarities can be easily found. Figure 5.4 demonstrates how these vectors and their relationships can be explored visually. The main downside of these vectors though is that they can not differentiate between homographs (words that are spelled the same but have different meanings e.g. river *bank* and money *bank*). Recently other models have been introduced that are able to reduce this problem by modelling the context of the word. The most promising of these models are BERT (Devlin, Chang, Lee, & Toutanova, 2018) and GPT-2 (Radford et al., 2019). These models are known as PTMs.

5.5 Pre-Trained Language Models

PTMs are a particular class of language models. They are also referred to as Large Language Models (LLMs) or foundational models. As mentioned in the introduction to the thesis and the introduction to this chapter, PTMs are different to the normal classes of models that have been introduced so far. The main difference with a PTM is that it has already been partially trained to understand language. That is PTMs are firstly *trained* to understand a language before they are *fine-tuned* on a specific task, for example classifying burglary MOs.

A useful analogy for understanding PTMs is a university student embarking on their first graduate job (PTM) compared to someone without education (generic ML model). The training to be successful at the job will have two parts. First, the trainee receives broad formal education, culminating in a university degree. They have a lot of knowledge and understand broad concepts, but it has taken many years and lots of effort to get them to that point. When they reach their new job, they will need additional, job-specific training tailored to the problems they need to solve for that role. The graduate needs additional domain knowledge, which will build on the broad concepts that they already understand. However, this additional knowledge is quicker to impart because of their already broad understanding of the underlying concepts. In the case of humans the cost of the education is somewhat dependant on the amount of people to be educated, however humans have a significant drawback that computers models do not have - they can not be easily replicated. As PTMs can easily be replicated (you can download a copy of one from the internet in minutes) that upfront training cost is only bourne once.

Using a PTM is like employing a graduate for the first time. The model already has some understanding, or knowledge, of the problem. In this case, the problem is what English words mean. However, the model does not understand the specific problem well. Therefore, the model must be given some on the job training before it is released for work. Previously, one could not "employ a graduate", one had to do all the model training oneself. Now, with the introduction of PTMs, one can "employ a graduate" and so skip most of the training. This means that significant complexity and effort in using NLP models has been removed from the end user. The first part of the training for PTMs is called pre-training and the second part is called fine-tuning. In this work only the fine-tuning will be conducted.

In addition to this two part training the PTMs also have a mechanism, called attention, that allows the model to understand context. Broadly attention allows the PTM to understand how the context of a word effects the meaning of another word in the sentence or the overall meaning. For instance the presence of "river" next to bank will lead the model to representing the sentence as a waterway rather than a financial institution. Likewise the presence of "not" near "good" is likely to lead to a more negative sentiment than a positive one.

PTMs are deep models, they are based on layers of neural networks that are trained to modify the input to output the correct result. As mentioned in Chapter 4, deep models are useful because they can reduce the need for feature engineering. Feature engineering highlights the most important features of a model input. The hard part of feature engineering is knowing which features to engineer and then finding a suitable representation. Examples of feature engineering were given above e.g. PoS tagging and NER. Feature engineering is time consuming and so by using deeper models the effort to establish a model is reduced. The downside to deep models however is that explaining why a model has made a decision is more difficult. Expainability techniques are therefore required to understand how and why deep learning models including PTMs are making decisions.

As PTMs offer the best performance across a range of NLP tasks they will be the model type used in this research. The exact PTMs and how they are built will be detailed in the methods chapter.

5.6 NLP Conclusions

This section has demonstrated that there are modern techniques available to assist with the extraction of information from unstructured free text data. These techniques have, for the most part, been developed into open-source models (PTMs) that can produce state of the art results on the material for which they were trained. These open models are coupled together into a processing pipeline, which relies on the success of each step to produce a numerical representation of the subject text that can then be explored through the use of the aforementioned ML techniques.

However, when these models are used outside of the types of data they were trained on, their efficiency drops. This is especially true if the underlying structure or grammar, of the text changes. As the focus of this research will be centred on police data, the next chapter will survey which and to what extent NLP techniques in this section have utilised with police generated free text data and what kind of benefits they produced. It will be shown that NLP techniques have been used to good effect on police data - although the practice is not widespread. The next chapter also highlights a gap in the research, in that PTMs have not been used with police free text data, and so their performance in this area is unknown.

Chapter 6

Natural Language Processing with Police Data

6.1 Literature Survey

Having demonstrated a need for enhanced analytical power for POP in Chapter 4 and the new techniques available for extracting information from text in Chapters 5 and 6, this chapter now maps the extent of the current research in the intersection of NLP and police generated free text data. This mapping is conducted through a literature survey, which shows that despite some utilisation of NLP techniques, there is a gap for the use of supervised learning techniques built on open-source models to extract pertinent information for crime prevention work. Additionally, few of the models found in the survey have demonstrated extrinsic utility – that is, utility for the ultimate stated purpose of crime prevention. Therefore, quantifying this extrinsic value is key to judging the importance of NLP techniques to POP and other crime prevention efforts.

Machine learning, text mining and data science have long been seen as useful tools for crime science (Marshall & Townsley, 2006). However, as a recent review into the intersection of crime and AI has shown (Campedelli, 2019), although some methods of AI and machine learning are relatively prevalent in the criminology literature, NLP and text mining are not that prevalent with relation to crime data. Neither NLP nor text mining get a mention in the top ten keywords of those articles discovered by Campedelli (2019).

Much of the crime free-text analysis is currently dominated either by non-supervised learning see - (Birks, Coleman, & Jackson, 2020; Kuang, Brantingham, & Bertozzi, 2017; Seo et al., 2018) - and revolves around the problem of crime linkage rather than crime reduction (Hassani, Huang, Silva, & Ghodsi, 2016). Recently however the complexities of the models have increased and there has been work to extract specific information directly from police free text data, (Karystianis et al., 2019; Karystianis et al., 2018). What follows is the results of a scoping review (Arksey & O'Malley, 2005) into the use of NLP with police generated free text data.

The scoping review was conducted with the aim of establishing *What is known from the existing literature about the utility and extent of Natural Language Processing with police generated free-text data.* Although quite a narrow search question, the previous two chapters have demonstrated that this research will nest into much larger bodies of work that are well established and documented. That is, the use of NLP techniques extend far beyond what will be discussed in this literature survey, and many of the techniques explored in the previous sections will be highly useful to this research to guide experimental design and model selection. However, as an emerging field, it is useful to understand exactly what has been achieved in the field of police generated free text data and NLP.

The literature review was conducted in four steps in accordance with Arksey and O'Malley (2005):

1. State research question. *What is known from the existing literature about the utility and extent of Natural Language Processing with police free-text data*

2. Identifying relevant studies. This was completed through searches of online databases. Scopus and Web of Science for journal articles and EThOS for Phd theses.

3. Study selection. Once identified the studies were read to ensure suitability. If found suitable then the references were checked for further studies.

4. Reporting the results. The results were synthesised and are reported

below.

6.1.1 Identifying relevant studies

The search for journal articles and proceedings was conducted through Scopus and Web of Science and the search of past PhD theses was conducted using EThOS (administered by the British library). The details of the searches and the number of items identified and found suitable are at Table 6.1. Although the search terms varied slightly between databases, essentially, they were all made of three components. The first was highlighting the need for a link to the police or crime literature. The second search component related to the analytical process of NLP and text mining, and the third component emphasised the focus on text data. In total, the database search found 38 unique and provisionally useful studies.

Database	Search Terms	Number of Items	Number of Relevant Items
EThOS	”Modus Operandi” OR ”police” OR ”crime” AND ”text” OR ”analysis” OR ”data mining”	122	1
Web of Science	AB = ((police or policing or crime) AND (”NLP” OR ”text mining” OR “information extraction” OR “entity extraction” OR “data mining” OR ”topic modeling” OR “classification) AND (text))	126	19
Scopus	ABS(((police OR policing OR crime) AND (”NLP” OR ”text mining” OR “information extraction” OR “entity extraction” OR “data mining” OR ”topic modeling” OR “classification) AND (text))) AND (LIMIT-TO (DOCTYPE , ”cp”) OR LIMIT-TO (DOCTYPE , ”ar”) OR LIMIT-TO (DOCTYPE , ”re”))	199	34

Table 6.1: Database survey search parameters.

Table 6.2: Literature Survey Results.

Authors	Purpose	Description	Intrinsic validation	Extrinsic validation	Notes
Birks et al. (2020)	Efficiently group crimes within a single crime classification with the aim of developing a greater understanding of crime problems and supporting design of crime reduction interventions.	Crimes from a single crime category, burglary, were analysed using Latent Dirichlet Allocation. Clusters were based on topic coherence. The clusters were then utilised to build a dashboard for analysts to geographically depict the distribution of the crime topics.	Number of clusters was based on maximising topic coherence score.	Dashboard app to geographically depict topics was built but no evaluation by practitioners of app or topics was reported.	Primary model = LDA Word representation = TF-IDF Feature engineering = n-gramm frequency analysis. Specific information Extraction = None Data origin = UK
Basilio, Pereira, and Brum (2019)	Identify demand structure in certain geographic areas to enable selection of appropriate policing technique.	Crime call transcripts are analysed as bags of words using LDA. Similar crimes are then grouped and ten topics for the two areas of interest were discovered. These 20 topics were then analysed to define the type of demand in each area.	Domain experts were used to validate and name the topics discovered.	None stated	Primary model = LDA Word representation = None Feature engineering = None Specific information Extraction = None Data origin = Brazil
Karystianis et al. (2019)	Automatically extract domestic violence information such as relationship and abuse types.	Rules based classification. 100 narratives were read and used to build sophisticated and extensive lexical rules. These were then validated against a further 100 narratives, and new rules introduced to account for any errors. These rules were then tested against a further 100 labelled narratives.	There was a test set of 100 narratives. Data had been labelled by domain experts. Precision of 90% for abuse type and 85% for victim injuries.	None stated	Primary model = Lexical rules Word representation = None Feature engineering = Rules and dictionary crafting. Specific information Extraction = DV types Data origin = Australia

Table 6.2: Literature Survey Results.

Authors	Purpose	Description	Intrinsic validation	Extrinsic validation	Notes
Krause and Busch (2019)	Analysis of accident type	Police narrative road accident data was analysed through a bag of words approach to classify into one of six different road accident types. Data that had already been labelled by the police was used as the training set, test data, which hadn't between labelled because it was not warranted, was used as the test set. Classification was done through SVM model.	69% accuracy across 6 classes on a held out test set.	None stated	Primary model = SVM Word representation = Bag of words Feature engineering = None Specific information Extraction = Classification only. Data origin = Germany
Haleem, Han, Harding, and Ellison (2019)	Strategic and Operational planning for mental health incidents	Classification of police narrative reports, into mental health related incidents or not. Police and academics were used to label the police narrative reports. Experimented with 5 different word embeddings and found a corpus generated Distributed Memory to be the best.	Cross validation only on the training set- 89.5% compared to the human generated labels.	None stated	Primary model = CNN Word representation = 5 varieties tested. Feature engineering = None Specific information Extraction = Classification only. Data origin = UK
Chohlas-Wood and Levine (2019)	To aid investigators in identifying groups of related crimes. (In this instance groups are related to the same serial offender, more crime linkage than POP.)	The whole model used a variety of variables including time and space. MO data was used as one of 30 variables. The comparison of MO data was through similarity score using bag of words model.	The narrative free text data was consistently ranked in top 5 of feature importance for the overall crime linkage model.	Deployed to every police officer in NYC PD.	The narrative was used as part of a much larger model. Primary model = Cosine similarity. Word representation = Bag of Words Feature engineering = None Specific information Extraction = None. Data origin = USA

Table 6.2: Literature Survey Results.

Authors	Purpose	Description	Intrinsic validation	Extrinsic validation	Notes
Karystianis et al. (2018)	Automatically identify type of mental health from police narrative data.	Rules based classification. 100 narratives were read and these were used to build sophisticated and extensive lexical rules. These were then validated against a further 100 narratives, and new rules introduced to account for any errors. These rules were then tested against a further 100 labelled narratives.	Test set of 100 held out examples. Precision of 97.5% of offender and 87.1% of victim disorders extracted.		Primary model = Lexical rules Word representation = None Feature engineering = Rules and dictionary crafting. Specific information Extraction = MH Conditions Data origin = Australia
Seo et al. (2018)	To automate the classification of crimes as either gang related or not.	The whole model was a Partially Generative Neural Network. The model used a variety of variables including weapon type and location information along side a narrative variable.	Narrative variable was included in the final model. Only 6 of 19 variables were selected.	When conducting sensitivity analysis around missing features, they found the narrative variable to be the most important.	The narrative was used as part of a much larger model. Primary model = Ave Word2Vec score. Word representation = Word2Vec Feature engineering = None Specific information Extraction = None. Data origin = USA
Pandey and Mohler (2018)	Exploration of metrics for crime topic modelling.	They cluster crimes from seven different administrative classifications into seven different topics. They then compute gini coefficients and topic coherence to compare with the data grouped and analysed in the original crime categories.	Spatial concentration was found to be higher in crimes grouped under topic models against those grouped by their administrative crime classification.	None stated	Primary model = LDA Word representation = TF-IDF Feature engineering = None Specific information Extraction = None Data origin = USA

Table 6.2: Literature Survey Results.

Authors	Purpose	Description	Intrinsic validation	Extrinsic validation	Notes
Kuang et al. (2017)	Goal is to discover ecologically more meaningful latent crime classes than the standard administrative classes, such that those crimes with similar conditions and or processes are grouped together.	Using MO data they use NMF, an unsupervised learning technique, to cluster the crimes according to their MO text. No other variables were used.	Cosine similarity between administrative classifications, generated through NMF process demonstrates expected similarities. Crime topics discovered the divide between property and violent crime reasonably well.	None stated	Primary model = NMF Word representation = TF-IDF Feature engineering = None Specific information Extraction = None Data origin = USA
Rogerson (2016)	Asses the feasibility of undertaking systematic analysis of descriptions of high volume crimes for crime prevention.	MO data was cleaned and prepared then bag of words models were used for k-means clustering. Analysis was based upon theft from person and robbery crimes.	Clusters validated against conceptual frameworks. Though clusters did not cluster around the same type of characteristic. For instance one clustered around an environment variable, while another clustered around the type of property stolen.	None stated	Primary model = K-means clustering Word representation = TF-IDF Feature engineering = Manual similarity engineering. Specific information Extraction = None Data origin = UK

Table 6.2: Literature Survey Results.

Authors	Purpose	Description	Intrinsic validation	Extrinsic validation	Notes
Helbich, Hagenauer, Leitner, and Edwards (2013)	Promote text mining, particularly the self-organizing map algorithm and its visualization capabilities, in combination with point pattern analysis, to explore the value of otherwise hidden information in a geographical context	Using text documents from a single criminal case they cluster using SOMs then plot the clusters on a normal geographical map to represent free text geographically.	None stated	Introduced new relationships between evidence items to the investigative team.	Primary model = SOM Word representation = TF-IDF Feature engineering = None Specific Information Extraction = None Data origin = USA
Bache, Crestani, Canter, and Youngs (2010)	Use language models to infer the characteristics of an offender. Individual crime data sets were labelled with characteristics and statistical tests were used to ascertain if the language models for each classification were different.	A variety of crime types were used to try and predict offender characteristics from the description of the crime by the police narrative data. These characteristics included age, sex and ethnicity. The narrative was represented by a bag of words model that was then analysed through Bernoulli models to generate word probabilities for each word in the respective classification. Offender characteristic classifications was then calculated through aggregating individual word probabilities from the respective crime narrative.	The models were only utilised on the training data. They find statistical significance for all crimes, but not all characteristics with each crime type.	None	Primary model = Bernoulli model Word representation = Bag of Words Feature engineering = None Specific information Extraction = None. Classification only. Data origin = UK

Table 6.2: Literature Survey Results.

Authors	Purpose	Description	Intrinsic validation	Extrinsic validation	Notes
Poelmans, Elzinga, Viaene, Van Hulle, and Dedene (2009)	Improve accuracy of Domestic Violence classification for violent crimes.	Using features from known domestic violence crimes they classify crimes into two groups. Using ESOM to analyse the features and nearest neighbour to classify they then use MO data to classify other crimes as either DV or not.	The classification accuracy was only 76%. However they were able to mitigate that by highlighting those cases with sufficient uncertainty for additional manual classification.	Companion studies show the model was used to further improve the definition of domestic violence and to improve police training.	Primary model = ESOM Word representation = Bag of words Feature engineering = N-gram frequency extraction Specific information Extraction = None Data origin = Netherlands
van de Putte, Oling, and Schakel (2009)	Search for intelligence from observational messages about suspicious situations, amongst police routine activity reports.	Using features of the police messages and the content, tree based rules were used to classify the messages as either administrative or as reporting suspicious activity.	The algorithm was only validated on the train data set. Accuracy was 85% though not reported explicitly.	None stated	Primary model = Rule based tree models. Word representation = Bag of Words Feature engineering = Properties of message such as length Specific information Extraction = Classification only. Data origin = Netherlands
Cocx and Kosters (2006)	To automate the process of crime linkage through determining a distance measure between similar crimes.	The whole model uses multiple document styles and is centred around cases rather than a single narrative. Entities are extracted from case documents, then similarity is measured between cases using matrix manipulation. Entities are extracted using SPSS software.	None stated.	In process of being validated by the respective investigative teams.	Primary model = Entity extraction Word representation = N/A Feature engineering = None Specific information Extraction = None. Data origin = Netherlands

6.1.2 Study selection

The studies from the searches above were investigated, and duplicates and unsuitable studies were removed. From 38 unique studies, 11 were found suitable on closer inspection. The selected studies were read and the references investigated to gather further suitable studies. In addition, local subject matter experts were consulted for additional references. At the end of the study selection, there were 16 suitable items of research. Most of the studies that were filtered out did not focus on police generated narrative data. Rather, they were focussed on news articles describing crimes. The studies selected, along with a brief overview, can be found in Table 6.2. Where the studies have only used NLP models as part of a larger model, the description focusses on the NLP element.

6.1.3 Reporting the results

During the investigation, no overarching review of the area in question was found. That is, there was no review into the utility of NLP and police generated free text data. Two reviews of a more general nature were identified (Hassani et al., 2016; Krishnamurthy & Kumar, 2012).These reviews were of a more general nature and did not concentrate on either police data or NLP. A good proportion of the studies from these two reviews included free text data that was non-police crime data, such as news reports, and so would not be subject to the same issues that MO data has, such as poor grammar and dialect (Keyvanpour, Javideh, & Ebrahimi, 2011). Neither review was systematic or identified specific search criteria. The selected studies generated from all searches were analysed, and the key observations from the studies are as follows:

Information extraction from police free text data is possible. There has been widespread evidence that useful information can be sourced from police free text data. The usage of police free text has been remarkable, with utility ranging from quantifying road accident black spots,(Krause & Busch, 2019) to drafting prosecutors indictment statements (Chen & Chi, 2010). Although most of the evidence comes from processing of the police narratives some older studies, demonstrate that MO data has been systematically used effectively for some time for crime prevention work Bowers and Johnson (2004) and (Adderley & Musgrove, 2003) albeit not using NLP.

One of the most intricate models has been used to produce a series of research exploring police free text data to uncover domestic violence and mental health issues in Australia (Hwang et al., 2020; Karystianis et al., 2019; Karystianis et al., 2018). Utilising an off the shelf NLP framework, General Architecture for Text Engineering (GATE), the researchers formulate (247) semantic rules utilising additional medicine and diagnosis dictionaries to label the data. An example of a rule is the following text would be in a document, *continued to X the victim* where X is an assault type from one of the dictionaries. They achieve results of up to 90% precision by only accessing 200 examples of data for training. They complete this work for mental health, domestic violence and an investigation into autism, extracting information that hitherto has been too time consuming to extract.

However, the effort to produce the rules and dictionaries is not reflected in the research, so the utility of this approach when dealing with changing information requirements is difficult to quantify. Another weakness of this approach is the changing nature of the terms used. These dictionaries and rules will need to be kept up to date with modern terms, such as new drug names, if they are to be used on a continuing basis. A further weakness that the authors identify is that similar but unknown terms are not picked up by the rules. This problem can be ameliorated by utilising word embeddings that allow similar words and phrases to be identified.

Rogerson (2016) is a thorough exposition of British MO data including an analysis of free text data, though they acknowledge none of the processes used were automated. Importantly the thesis demonstrates that there is information in the MO data that is useful for crime prevention work, though that the crimes analysed do not fit neatly and exclusively into the crime codes given, drawing similar conclusions to that of Birks et al. (2020) and Kuang et al. (2017), that administrative codes hide crime variation.

Utility of NLP and crime data is not limited to Modus Operandi data. Helbich et al. (2013) demonstrate that NLP techniques can be used across a variety of documents, within one investigation, and the results drawn together to produce “useful” insights. Though due to the sensitivity of the case what the insights were or how effective they were can not be divulged. A separate studies (Cocx & Kosters, 2006), focussed on crime linkage has shown that models can be used to identify links between crime incidents. This was achieved by reviewing different documents from the same case, though there is little practical evidence presented that the links have a real world practicality.

Most of work so far has been unsupervised learning. Notable examples of this are Birks et al. (2020) and Kuang et al. (2017) who use unsupervised NLP to understand how crimes may be grouped relative to how they were committed rather than traditional crime classifications. Birks et al. (2020) completes this within a crime classification and Kuang et al. (2017) conducted this across multiple crime classifications. This is referred to a crime topic modelling and seeks to understand crime from an ecological perspective. This idea is extended further by Pandey and Mohler (2018) who investigate the crime topics through spatial distribution, suggesting that as crime is also a function of an environment then the spatial concentration of a crime topic can be seen as a proxy measure for its coherence. In relation to problem solving they were also able to successfully group sub-categories of crimes using clustering techniques, which could then be used to identify interventions. However the clusters did not partition along the same characteristic of the crimes, some clusters were partitioned on the type of environment and some on the type of objects involved in the crime. This means that only partial information about each crime is being used to cluster the crimes, as the topics are predicated on only the most likely words.

In addition to the previous studies there have been a pair of studies conducted with police data from Brazil, (Basilio, Brum, & Pereira, 2020; Basilio et al., 2019), that have used unsupervised NLP techniques to cluster crimes to begin to understand what policing strategies will be suited to different areas of the city. They clustered the crimes, then showed police officers a representative sample of the clusters to name a suitable policing style (traditional, POP etc). They do not report if the styles were subsequently adopted or if they were successful.

Unsupervised learning has presumably been popular because it can be conducted in a computer lab, with minimal resources and/or input form practitioners. The unsupervised research is very much exploratory, but as of yet this research has yet to prove that those results found have utility for crime prevention. Kuang et al. (2017) investigate their results and prove that they have found partitions along violent and property crime, and separately between gun and non-gun crime, though presumably they would have been sorely disappointed had these divisions been missing. Birks et al. (2020) investigate their results through presentation of a dashboard however this, nor their topics, are validated by practitioners as a useful crime prevention tool. Pandey and Mohler (2018) utilise the idea of spatial coherence to strengthen the validity of their crime topics, but fail to account for environmental descriptors or entities in the data or the fact that police officers in the same areas may use similar language. Additionally most of the unsupervised learning has yet to explore more powerful methods of word embedding that may have strengthened their results, word embeddings may have been able to link words of similar meaning and thus overcome some of the choice of language that may be unnecessarily partitioning the topics.

Prevalence of classification The results also show that classification of incidents has been more prevalent than specific information extraction. As noted in the earlier chapters, particularly Chapter 3, specific information about an incident is required to group similar incidents with similar processes. Classification of incidents is useful, and the research found has shown that classification of incidents is possible (see next paragraph for evidence). However, the focus on classification means that actual details from the text have not been extracted. For example, a classification technique will classify if force has been used in a burglary or not, whereas a more sophisticated technique for information extraction may extract the type of force. For example, "smashed window" or "jemmied door" will be extracted, instead of just being classified as force used.

As an example of classification with extrinsic validation, police free text data was used to better classify incidences of domestic violence that had previously relied on officers tagging keywords. Utilising a machine learning technique, Self Organising Maps, (Poelmans, Elzinga, Viaene, Van Hulle, & Dedene, 2009) were able to more accurately label those incidents that included domestic violence, and with that information, they were able to better educate officers to recognise domestic violence and also help to better define the issue. (Poelmans, Elzinga, Viaene, & Dedene, 2009; Poelmans, Elzinga, Viaene, Van Hulle, & Dedene, 2009; Poelmans, Van Hulle, Viaene, Elzinga, & Dedene, 2011).

Seo et al. (2018) takes a slightly different approach to classification by trying to understand which crimes are gang-related. They use free text description as a narrative variable among a host of other structured variables to feed a neural network. They find that of all the variables used, the narrative ones were the most important for the model – highlighting the valuable information contained in the narrative reports. However, as they only use an average of the documents' word vectors, most of the information in those documents will have been lost.

Bache et al. (2010) use free text MO data (and keywords) to try and predict offender characteristics such as ethnicity and employment status. This was achieved through a bag of words approach, then a form of reverse topic modelling with known topics, such as male or female. Once split into these known topics, the defining words in the topics were used to understand the characteristic unique to those topics.

Where supervised learning has been used feature engineering was key to success. This was especially true using shallow models (i.e, nonneural networks), as one would expect. This serves to further highlight the trade-off that utilising machine learning will bring to police analysts. Shallower models will require more input, and possibly longer to build as the features are developed; however, they may offer greater insight into why classifications were labelled. (Bachenko, Fitzpatrick, & Schonwetter, 2008; Ku & Leroy, 2013; van de Putte et al., 2009). It is possible to partially automate feature engineering through the use of neural networks; however, as explained above, this may lead to a reduction in the explainability of the model.

Corpus generated word embeddings work better. Where word embeddings have been mentioned, they have indicated that embeddings generated from the data themselves have been better than pre-trained models such as Word2Vec (Haleem et al., 2019; Schraagen & Bex, 2019). This again reflects on the difference between police data and the more widely used (often edited) data that is traditionally used for pre-training open-source models. This evidence reinforces the need to ascertain how effective PTMs are and how they can be tuned if necessary.

6.1.4 Police and Algorithms

Although not specifically associated with NLP, Babuta, Oswald, and Rinik (2018) is a RUSI publication that explores the use of algorithms in a UK police context. However, the paper's focus is on predictions of individuals' proclivity for future crime, rather than crime events themselves. The paper highlights the lack of frameworks and direction from central policy makers in algorithmic usage for UK police forces. However, there is one framework that has been partly adopted by the National Police Chiefs Council that is currently filling the policy void. This framework is ALGO-CARE.

ALGO-CARE

One tool that has been developed and partially adopted by the a police governing body for UK police (National Police Chiefs Council) is ALGO-CARE (Oswald, Grace, Urwin, & Barnes, 2018) . ALGO-CARE was developed alongside an automatic risk assessment tool in Durham police force and is a "decision-making guidance framework for the deployment of algorithmic assessment tools in the policing context" (Oswald et al., 2018). In short, ALGO-CARE is a mnemonic that has been developed to allow police leadership to understand whether or not to deploy an algorithmic tool. The mnemonic is explained below.

- Advisory. Is there a human in the loop? Or is the process or tool fully automated between input and output.
- Lawful. Is the purpose necessary and legitimate for policing purposes?
- Granularity. This factor encompasses granularity of data and decisions at all levels and is split into 6 sub-areas.
- Ownership. Who owns the algorithm and the data on which it is trained?
- Challengeable. What are the post-implementation oversight and audit mechanisms to identify any bias?
- Accuracy. Covers all elements of performance fo the algorithm. Essentially is the performance good enough for the intended usage.

- Responsible.Covering such elements as a fair, accountable and ethical approach.
- Explainable. Can appropriate information be given about how the algorithm has come about each score?

Although ALGO-CARE was developed for risk assessment tools, it has wider applicability. The most important message that the tool conveys is that good performance (their Accuracy) is not the only consideration for implementation. Other requirements, such as on what problem the tool is used, how officers use the information and issues of fairness are all important factors for the utilisation of modern NLP models. This spread of requirements is reflected in the Methods chapter, where the performance of the PTMs is not just predicated on a measure of correctness (the metric used will be MCC) but also explainability and bias.

6.2 NLP with Crime Conclusions

In summary, there has been a spread of use of NLP with police generated free text or narrative data. Most of that presented in the literature survey has been intrinsically successful; that is, the models built have generally been found to have accurate results, giving confidence that using NLP with police free text data is possible. Extrinsic validation has been less well shown, however, meaning that the models built have not shown real utility for their intended ultimate purpose. This, in part, may be due to most of the research emanating from the computer science community, which may not benefit from the stronger working relationships with police forces that the criminologists have. The NLP success rate will of course have benefited from publication bias, but knowing the relative recent success of NLP in the larger community, and against published standardised data sets, it is not surprising that success has been found with police free-text data.

Within the literature survey, no examples were found of models built using PTMs. PTMs are the more modern style of models that were introduced in Chapter 5. PTMs have already been partially trained on the English language and just need additional fine-tuning on the NLP task required. As mentioned in Chapter 5, PTMs can be more accessible to potential users as they require less feature engineering and so can be used with less technical knowledge. The focus in this thesis is on understanding if PTMs can be used to generate information from police free text data and thus offer insight into this research gap.

Given the additional burden required for labelling to enable supervised learning, it is not surprising that most of the work has focused on unsupervised techniques. However, when labelled data has been provided, it is clear to see that more specific information has been extracted, indicating that labelling data for police free-text is a worthwhile endeavour, and any activity that can reduce this burden is going to have real practical significance.

The existing research demonstrates that useful information can be extracted from police free-text data. The data extracted can have utility for crime prevention strategies, but the practical application of NLP systems has not been fully tested. However, particularly for UK police free-text data, there is no example of an automated NLP solution for information extraction that has proven portability across different crime types and police forces.

The next chapter restates and explores the research questions that will be answered in this thesis. This will be the final chapter of this part of the thesis.

Chapter 7

Aims and Objectives

This chapter is the final chapter in the first part of the thesis. The main aim of this chapter is to state the research question and the sub questions to set the agenda for the rest of the thesis. Firstly the main research question and the sub-questions will be stated. Then the main research question and the sub-questions are individually explored. Finally a table is given to show where each research question will be answered related to the studies in part 2 of the thesis.

7.1 Primary Research Aim

The thesis will be motivated by the following overarching research question:

Can PTMs be used efficiently to extract information from police free-text data, and if so what practical applications for problem-oriented policing does this approach have?

With supporting objectives:

- Identify the extent of NLP usage with police data.
- Evaluate how effective PTMs are with MO data.
- Evaluate how effective PTMs are with police incident data.

- Evaluate how effective active learning is with police data.
- Identify which parts of the POP process might be best supported by the use of PTMs.
- Identify implementation barriers for PTMs.

The next section explores each of these research questions.

7.2 Primary Research Question

Can PTMs be used to extract information from police free-text data, and if so what practical applications for problem-oriented policing does this approach have?

7.2.1 Aim

The overarching aim of this study is to extend research into the effectiveness of NLP methods at analysing police free text data. The focus will be on how modern PTMs can be used to classify police texts. This classification will enable a greater understanding of intra-crime variation and so may provide insights into the specificity of problems, thus supporting the application of POP.

7.2.2 Discussion

The previous chapters have discussed how POP focuses on specificity in identifying and understanding problems. Subsequently, the challenges associated with conducting the POP process were discussed. It was highlighted that free text data, while rich in content, was often underutilised in this process due to a range of logistical challenges.

Chapter 5 discussed how a diverse range of NLP methods can be used to extract meaningful insights from free text data and how fruitful applications of these methods have been observed in a range of domains. More recently, NLP techniques have developed to produce a class of models known as pretrained language models (PTMs). PTMs can analyse texts with little additional feature engineering. These models have proven themselves useful with established academic test sets, but they have not yet been knowingly tested against police generated data. If the PTMs are able to effectively analyse police generated data, it is likely that they will be able to alleviate some of the time-consuming analytical processes in POP.

When considering if PTMs can be used to extract information from police free text data, there are wider concerns than just the accuracy of the model. Therefore, other factors that allow a technique, especially an AI technique, to be used in a public service capacity will also be considered. The additional factors will be model bias and explainability. These factors were introduced at the end of Chapter 6 and were seen as important to track to understand if a model was suitable for use.

The next sections introduces the supporting objectives, which will be explored to answer the main research question.

7.3 Supporting Objectives

7.3.1 Identify the extent of NLP usage with police data.

This was already completed in the previous chapter by way of a literature survey. As a brief recap, it was found that NLP techniques have been used with police generated data, both academically and practically. However, the NLP work is not extensive, and no work with PTMs has been recorded. This sets the research objective of reviewing the performance of PTMs with police generated data to understand how these PTMs can be used with little additional manipulation and what results they will offer.

The next two objectives focus on the use of PTMs with two different text types. The two different types are MO data and police incident logs. These data types were briefly introduced in the Introduction and will be described more fully in the Data chapter.

7.3.2 Evaluate how effective PTMs are with MO data.

MO data is a short description of a crime. MO data records aspects of intra-crime variation for each crime and so can be a useful source of information for POP practitioners. This objective will investigate how well PTM models can extract this intra-crime variation. The evaluation of this objective will be conducted in Study 1. Study 1 will focus on MO data, in particular burglary MO data. The data for this study will be drawn from two police forces known as PF1 and PF2. These police forces will be described in the data chapter.

Model effectiveness will encompass performance, explainability and the presence (or absence) of bias. In addition, where data availability allows, performance over time and across police forces will be investigated. Performance over time is important because effort spent building models that last longer will be more resource efficient. Performance across police forces is important because wider use of a single model means more efficient use of the resources required to build the model.

7.3.3 Evaluate how effective PTMs are with police incident data.

Police MO data is not the only data that describes police problems. Another ubiquitous source of data is police incident logs. For this reason, PTM effectiveness will also be judged against police incident logs, particularly police incident logs describing anti-social behaviour (ASB). The police incident log data is only drawn from PF2. Choosing another text type is important because different texts are formed in different ways and can use different language. These linguistic differences may mean that PTM effectiveness changes between text types.

7.3.4 Evaluate how effective active learning is with police data.

Using PTMs generally requires labelled data. Labelled data, however, requires resources to generate, resources which would otherwise be applied to POP problem solving in other ways. One method that has been developed to reduce this resource requirement is active learning. In keeping with lowering the burden on POP, active learning will be investigated to understand what kind of efficiencies can be achieved by adopting the technique.

7.3.5 Identify which parts of the POP process might be best supported by the use of PTMs.

The proceeding three supporting objectives will form the middle part of the thesis. This final supporting objective will review what has been learned in that middle part of the thesis and then suggest how the POP burden may be lowered. This analysis will be achieved using the SARA framework. The SARA framework is a framework for implementing POP that was introduced in Chapter 3.

7.3.6 Identify implementation barriers for PTMs.

Introducing any new practice or software is likely to hit barriers. These barriers can stop a new practice being implemented if they are not identified and addressed. This supporting objective highlights the most important barriers, suggesting how they may be overcome.

7.4 Conclusion

This chapter has set out the research question and sub questions that the studies in this thesis will go onto address. The next part of this thesis relates to the practical aspects of this thesis. Principally testing PTMs to see if they perform well with police free-text data. After Part 2 the third and final part of the thesis will draw together the results from Part 2 and use them to explore how PTMs might be best used in assisting POP interventions.

Data	Study Focus	Task details
Study 1a - Supporting Objective 2		
MO (PF1)	Information extraction	Classification - Is force used? Classification - Is a car stolen?
Study 1b - Supporting Objective 4		
MO (PF1)	Active Learning	Comparison of model metrics - Active learning v random selection
Study 1c - Supporting Objective 2		
MO (PF2)	Information extraction	Classification - Is force used? Classification - Is a car stolen? Classification - Outbuilding only?
	Transfer learning	Comparison of model metrics - Over time Comparison of model metrics - Across police forces
Study 2 - Supporting Objective 3		
Logs (PF2)	Information extraction	Classification - Traditional ASB Classification - Covid complaint Classification - Group present

Table 7.1: Study focus and objectives. This table breaks down what the focus of each study within this thesis are and what data is used.

Part II

Case Studies

Chapter 8

Data and Data Processing

8.1 Introduction

This chapter will introduce the data that have been used in the studies within this thesis. This chapter will give background information on the text data that is used with the language models, and how the composition of the data may effect the performance of the models that are utilised. The data was from two police forces, 1) PF1 and 2) PF2. Both police forces are located in the North of England.

The PF1 data underwent screening by PF1 before being released to The University of Leeds for a number of projects. PF2 data was primarily provided to the University of Leeds for the previously mentioned Covid-19 ESRC funded project. The author contributed extensively to screening of the PF2 data before it was provided to the University.

As both data sources went through some form of screening to remove personally identifiable information (de-identification) the data is not in the exact same form as the police services would use it. Redaction may have a negative impact on model accuracy, as information will have been removed, however it does mean that if police forces were to further utilise the PTMs they should expect better model performance as they will have access to the raw data.

8.1.1 Chapter Outline

This chapter will first introduce police textual data before introducing the datasets from PF1 and PF2 . The Chapter will then explain in greater detail how the PF2 data were prepared and desensitised for analysis away from PF2 servers. This preparatory step is of interest because free text data is highly likely to include personal data and so if researchers want to utilise the data away from police servers then they will have to implement steps to reduce the risk of personal data loss. Removing personal data references from the police free text data also allowed the use of non-vetted personnel for data labelling, a time consuming and laborious task, that nevertheless is critical for supervised machine learning tasks. The use of non-vetted personnel allows for much more flexibility in recruitment of the data labellers, and so greatly reduced the data labelling burden.

8.2 Police Data

The data used in all studies were exclusively police generated data. In particular the data was also sensitive police data, in that it originally held personal information and so is not freely available to the public. The use of sensitive police data generates two main problems. Firstly from a practicality perspective the data had to be de-identified, as mentioned above. Secondly, and more generally, as police data does not accurately reflect the totality of crime that is committed (Tarling & Morris, 2010a), this will have implications of bias, as introduced in the earlier NLP chapter. The next section introduces the two types of police text data used in this study, MO texts and incident logs. Weakness with the police data are returned to at the end of the chapter.

8.2.1 Modus Operandi Data

An MO text is usually a short text document of one to three sentences that describes the main elements of a crime. The MO is but one element of data recorded about a crime. MO data is not explicitly generated for crime prevention work. MO data is designed to be a short description of a crime that can be released to other agencies, typically still within the criminal justice system. This influences the content of the MO data, such that it should not contain personally identifying information or excessive details such as lists of stolen items. MO data makes up only a small portion of the data, including the free text data, that is recorded about an individual crime. Underneath the MO data sits a more detailled incident summary that contains more information as well as more detailed incident descriptions from witnesses. Examples of MOs can be seen in Table 8.1.

The selection of MO data had two benefits. Firstly the text passages were a relatively short but condensed description of the crime - as text passages become long they are much more computationally expensive to compute. Of course the trade-off with short text is that it can lack details about the crimes they describe. The second and perhaps more important was the lack of personal data in the text, this gave the police forces more confidence to share the text with us. Undoubtedly other sources of textual information, incident summaries and witness statements for example, will have more information to extract, but they are also riskier to share as they are likely to contain more identifiable data. MO data was therefore a pragmatic compromise between data security and data utility.

Typically the MO data for each crime is also accompanied by flags that help to explain intra-crime variation. Intra-crime variation here means the variation between crimes of the same administrative designation. As an example residential burglary is an administrative crime classification, but within that crime type there is variation such as the use of force or not to enter the property. Flags help to systematically (i.e. not in free-text) record intra-crime variation that is not otherwise recorded in the mandatory recording fields. Typically flags are an additional field that the police officers select to record specific details about a crime, such as the entry point of a burglary, or the use of a weapon in an assault. They are the digital equivalent of a check box at the end of a form. As the fields are not mandatory the completion rates can be poor, and in the studies within this thesis we are able to compare the NLP models to the Officer generated flags giving an indication of completion rate.

In summary, MO texts are short descriptions of the means by which crimes were undertaken. Generally, this included the known key events of a crime. They were designed to give a coherent overview of the crime. They may contain identifiable information, and quite often they are complemented by a series of flags that give further systematic detail on intra-crime variation.

MO 1	Attacked property is a privately owned end terrece multi occupancey dwelling. Between times stated suspect/s enter through insecure ground floor window. Tidy search conducted and vehicle keys removed from kitchen hooks. Suspect make their escape through same and leave stealing vehicles. Vehicle XXXXX found burnt out
MO 2	Modus operandi summary.....Attacked property is a mid-terraced property located on a quiet residential street. Between times stated unknown suspect approaches the front of the property and with bodily force kicks open the basement window. Suspects gain entry to the property and untidy search in conducted. Suspects exit property with stolen items and make off in unknown direction

Table 8.1: Two example MOs from the PF1 data, complete with errors. Reproduced from Birks et al 2020

8.2.2 Incident data

Incident data is collected by police on all issues that are reported centrally. Typically these reports are made by members of the public verbally through the use of emergency and non-emergency phone numbers to a central call station. However, they are also increasingly made using other messaging techniques such as email and online reporting tools. Examples of police incident logs are at Table 8.2

Police incident logs are generated as the information is received. They are the first record of an incident and they may or may not include a crime. For the purposes of this study the textual log data received only included incidents that were classified as anti-social behaviour (ASB - described later), so they were not designated as crimes. Logs can include the initial report, the first interactions of Officers as they attend the scene and subsequent reports. These subsequent reports can contradict the original report or add explanatory detail. Generally the logs are not edited, or rationalised to depict a single coherent narrative. This can make comprehension of a log difficult. For instance a report may be made that a Covid-19 rule was broken, but subsequent reporting from police officers may confirm that no rules were broken.

As demonstrated below logs are generally longer and have more word variation than MO texts. They also do not typically come with additional flags to help systematically record intra-incident variation.

8.3 PF1 Data

The PF1 data comprises crimes committed in a police force in the North of England. Two years worth of crimes were provided. Though the years were not specified. Crimes of a sexual nature and or related to domestic abuse were also withheld. All of the data fields supplied with the data can be found in Table 8.3. The MO texts came from the *Crime Notes* column and the flags came from the *MO Description column.* The PF1 data went through processes unknown to redact identifiable information from the MO texts before it was given to the University of Leeds. Only the Burglary crimes from the PF1 data were used for this study (the reason for this is explained at the beginning of Study 1a). The burglary MO texts are described next.

8.3.1 Burglary MO Data

The PF1 data contained 9818 burglaries. As mentioned previously the year of the crimes was not given but the day of the week and the month was given (Table 8.3). The median number of words in a MO text is 65, with the interquartile range being (48,88), see Table 8.6 for a comparison with the PF2 data. The longest MO was 403 words long.

As the main PTM to be used was BERT, it is also worth exploring if BERT will recognise the words used in the text. BERT can only recognise certain words or tokens. If the words are not recognised then they are broken into word pieces that are then recognised, although they may not have the same meaning as the original word.

Comparing the MO words with the BERT vocabulary shows that BERT recognises 96% of the MO words (by volume). The remaining 4% of words are broken into word pieces which the BERT model recognises. As BERT is trained on books and Wikipedia text, not police records, it is worth exploring which of the words in the MO text that BERT does not recognise. Table 8.7, shows the top ten unrecognised words. The table shows that there are unrecognised words, for example ‘insecure’, that may have an important bearing on describing the

Incident text 1	brother is throwing bricks at the window xxxxx he has mh issues - he is called xxxxx xxxxx xxxxx this has happened after an argument xxxxx xxxxx is outside the property now xxxxx xxxxx is shouting outside the house xxxxx no damage caused at the moment but is now throwing stones at the top floor flat given out xxxxx dob - xxxxx last name xxxxx xxxxx first name xxxxx xxxxx xxxxx birth date xxxxx relation type xxxxx 06 crime intelligence xxxxx xxxxx has anger management xxxxx . house is locked and secure xxxxx xxxxx xxxxx desc - white male , medium build , 5 ft , 9 , xxxxx brown hair , dark blue jacket xxxxx , light grey pants xxxxx still screaming xxxxx xxxxx symptoms of covid or xxxxx in xxxxx xxxxx xxxxx xxxxx had left prior to our arrival . there is no damage and no trace of him . no reports . cdit review - no ammendments to log as no offences disclosed .
Incident text 2	an email request has been made . default email notification has been made to xxxxx xxxxx . com . email received in fcm 22/10/2020 at 07 xxxxx 36 reference number xxxxx xxxxx incident relates to xxxxx individual location address xxxxx 1 xxxxx xxxxx street name of persons involved if known xxxxx xxxxx and her son is the subject displaying any covid 19 symptoms xxxxx yes time of incident xxxxx 07 xxxxx 30 date of incident xxxxx additional information xxxxx its every weekend now she is constantly breaking the rules but it doesn't matter her because she doesn't work anyway she's a xxxxx xxxxx and its really not fair now and she goes mixing with household with her sons it needs to stop but she won't listen and has been told by neighbours please cross refer into op talla master log log can be closed with thanks further email from the INFORMANT - 15 xxxxx hi that's fine thanks , please could you not mention any names as i don't want is causing any problems thanks

Table 8.2: Two example incident texts from the PF2 data. Note "xxxxx" are redacted words.

Crime Data Fields		
URN	Crime Type	OccType
Day	Month	PartialPostCode
MODescription	CrimeNotes*	HOClass
OffenceRec	DomViol	

Table 8.3: A table of all the data fields for the PF1 crime data.* Indicates a free text field.

crime. Although they will be broken into word pieces and not removed, this disassembling of the word may be a source of error that prevents the BERT language models from classifying the texts correctly.

8.4 PF2 Data

The second source of data was from PF2. PF2 is also a police force in the North of England. PF1 borders PF2. The main difference between the PF1 data and the PF2 data is that the PF2 data included crime data and police incident data. The transfer of data from PF2 to the University of Leeds was also more closely controlled by the author. The data extract specifications was built alongside the PF2 police analysts and the data was extracted as a joint effort. In addition the author built the de-identification process, described in detail in the next section, that was used to de-identify the free-text data. The PF2 data contained both structured and unstructured data fields. The initial data was transferred in January 2021, followed by a secondary data transfer, in February 2022 that allowed additional fields to be extracted for model verification purposes.

All crime data fields are shown in Table 8.4, incident data fields are shown in Table 8.5. The review of the data in this section will focus on two data types. PF2 burglary MO data, which was used to replicate analysis of the PF1 data used in study 1 and secondly anti-social behaviour (ASB) police incident logs that was used to investigate the use of PTMs on police incident logs. Each data type is now reviewed in more detail.

8.4.1 PF2 Burglary MO Data

The PF2 burglary data consisted of just over twelve thousand reported crimes. It includes all residential burglaries and attempted residential burglaries committed from 1st January 2018 to 31 December 2020. Each reported crime contained an MO text. The median number of words for an MO text is 31 (IQR 22,46). Comparing the PF2 burglary data to the PF1 data we find that it is generally shorter and more homogeneous, see Table 8.6, so possibly less descriptive. We would therefore expect models to be poorer as there is less variation in the data on which to discriminate.

After the modelling was complete PF2 released additional data to help quantify the effectiveness of the classification model built to identify when a car was also stolen. PF2 provided the results from a data search that showed when a vehicle had been linked to a burglary, and the link of association was “stolen”. Typically they expect this field to be more complete than text references in the MO data to a stolen vehicle - so it can not be used as a direct metric as the language models can only analyse information stored in the free text data. That is given the selective nature of free-text data a car maybe stolen and logged as linked to the burglary but not mentioned in the free-text MO description, and thus the information is not available to extract. For a complete list of fields provided see Table 8.4.

As with the PF1 data PF2 MOs were explored to see what percentage of words are contained within the BERT model vocabulary. By comparing the MO text with the BERT word list it can be seen that BERT recognised 96% of the words (by volume), the remaining 4% of words are broken into word pieces which the BERT model recognised. This is the same percentage as PF1. Table 8.7 shows the top ten unrecognised words. The top ten unrecognised does vary across police forces, although there is some overlap in meaning.

8.4.2 PF2 ASB Incident Logs

As described earlier incident logs are different to MO data in that they are generated primarily through reports made by members of the public. Incidents do not have to be crimes, and indeed the incidents that text data was received for were not classified as crimes. The incident logs had all been classified as

Crime Data Fields		
Investigation number	Call origin	Crown victim
Storm ref	Outcome	Town
Committed from date	Status	Postcode
Committed from time	Offence	Easting
Committed to date	Primary offence	Northing
Committed to time	Included offences	MO keywords
Reported date	Victim age	Factors
Reported time	Gender	MO Text*
Recorded date	Occupation	Relationship type†
Recorded time	Ethnicity	Linked vehicle†

Table 8.4: A table of all the data fields exported from PF2 for the crime data.* Indicates a free text field. † Indicates a field sent post analysis.

anti-social behaviour. Incident logs are typically much longer than MO data, as can be seen from Table 8.6 the median words in a document is over fives times greater than that of the burglary data standing at 166 for an ASB incident log. The inter-quartile range of word counts is also much larger at (100-290).

Table 8.7 shows the most common words from the ASB documents that are not recognised by the BERT model. Most of the words are abbreviations. Of note here is that "covid" is not recognised, this is because when the BERT model was trained (2018) covid was not quite as infamous as it is now. In total 11.1% of words in the incident logs are not recognised in their complete form by the BERT model. This is higher than the MO data, but not unexpected as the ASB log uses more place names, abbreviations and telephone numbers. The incident data has a smaller subset of data fields than the MO data, the fields provided are listed in Table 8.5.

Incident Data Fields		
Incident number	Complainant	Disposed time
Initial theme	Priority	Postcode
Final Theme	Origin	Easting
Initial	Input date	Northing
Final	Input time	Covid_19
External ref	Disposed date	Incident text*

Table 8.5: A table of all the data fields exported from PF2 for the Incident data.* Indicates a free text field.

8.5 Data cleaning and de-identification

This sections sets out the steps for the de-indentification the PF2 data. This was required to meet data protection requirements and took place before the data could be transferred from the PF2 servers to the secure University of Leeds servers.

Whitelisting was used as the method to de-identify the data. De-identification is the process of removing personally identifying information from the data. For structured data this is generally a trivial task, for example removing the second half of a postcode generalises the data sufficiently such that individuals can not be identified even in sparsely populated areas. Free text however is different. For the police staff who input text there are essentially no limits on what information can be included. Full names, addresses, date of births can all be entered into a free text box without technical issue, even though procedurally they should not be entered. In some instances, for example the incident data, personal information is expected and a necessary part of the data being logged. However, MO data is designed from the outset to be released to third parties, though principally still within the criminal justice system, and so should not routinely contain personal identifying information.

Medical research has studied the issue of de-indentifying data extensively. Models built for this task range from simple rule based models to more

	Median Words per document	IQR Words per document
PF1 Burglary MO	65	(48,88)
PF2 Burglary MO	31	(22,46)
PF2 ASB Logs	166	(100,290)

Table 8.6: This table contains descriptive statistics on the different text corpus used in the thesis. Median was used as the average as the distribution of words is skewed. IQR stands for inter quartile range and is the 25th and 75th percentiles.

intricate NLP based models (Meystre, Friedlin, South, Shen, & Samore, 2010). However there is no consensus that these models work perfectly in all situations (Narayanan & Felten, 2014). Each de-identification model style has downsides, the machine learning models require a lot of labelled data to train, and tend to be difficult to explain. The rule based models require extensive knowledge of the data and are not robust against unseen phrases within the data. For this research the most important characteristics for the de-identification process where easily explainable rules (to be explained to and understood by police staff) and a risk adverse approach (to avoid any data protection issues arising). For this reason a whitelisting approach was chosen.

Whitelisting is a well known, conceptually straightforward and safe approach. Whitelist methods uses a list of safe words. If a word in the police free text data is on the safe list then it is kept, if it is not on the list it is redacted. The resulting text is therefore only constructed from the words on the safe list. This is a simple de-identification method, which is easy to explain and deterministic, but has the downside of potentially redacting rare but important words.

The next section will explain the whitelist procedure in more detail. Data cleaning was used alongside this process to homogenise the text and improve the retention rate. Data cleaning is explained in the next section before the de-identification process in detail.

Data Set	Top Ten Non-BERT Words
PF1 Burglary MO	egress, insecure, untidy, complainant, upvc, terraced, semi-detactched, occupant, jemmy, comp
PF2 Burglary MO	XXXXX, undetected, insecure, untidy, aggrieved, terraced, burglary, trespasser, UPVC, unoccupied
PF2 ASB Incident text	XXXXX , inf , covid, npt , cctv , nuisance , informants , pls , pcso , fcm

Table 8.7: This table contains words that are not in the BERT language model list but are in the police data used. Only the top ten missing words by volume are listed. The words that do not appear in that list will be broken down into word pieces and so meaning may be lost. XXXXX is the symbol for redacted words.

8.5.1 Data Cleaning

In addition to screening the data for de-identification there was also data cleaning to homogenise the text so that information was not lost when certain words or tokens were removed. This included spelling correction, replacing jargon and replacing detailed information with a representative placeholder. The different aspects of this process are discussed below:

1. Misspellings. Common misspelling were identified. These were added to a misspellings list. This misspelling list was then used to correct words in the MO text before it was de-identified. There were just over 900 common misspellings and typography's identified that were then corrected. Misspellings were identified by hand.
2. Jargon. Although jargon would be identified through other processes later, some of the words or phrase were changed to represent words that were more likely to be recognised by the PTMs. An example was changing m/v to motor-vehicle. Again this list was generated through reading representative samples of texts.
3. Placeholders. Some information was replaced with generic placeholders denoting the type of information while removing personal identifiers. An

example was UK vehicle license plates that follow known formats which were replaced with the token 'NUMBER_PLATE'.

8.5.2 De-identification Process Overview

The de-indentification process used a white list approach. Words were removed from the text if they were not on the list of approved words, referred to here as the safe list. The safe list was built in an iterative manner. It was seeded with a list of frequently used English words (detailed further below) that were compared to the text. Those words not on the safe list were arranged in frequency order (most common at the top). This list was then reviewed and words deemed safe were added to the safe list. Common misspellings were identified and added to a data cleaning list. The process is shown in Figure 8.1. The outputs of the process were a list of safe words plus a list of transformations for common misspellings or abbreviations.

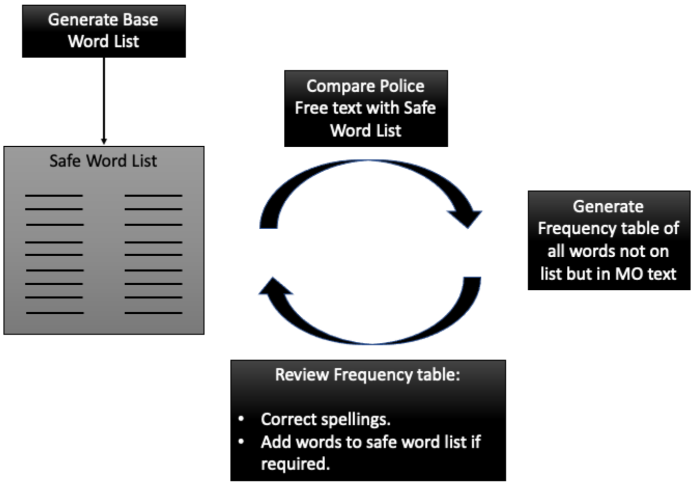

Generate Base Word List
Safe Word List
Compare Police Free text with Safe Word List
Generate Frequency table of all words not on list but in MO text
Review Frequency table:
• Correct spellings.
• Add words to safe word list if required.

Figure 8.1: This depicts the cycle to generate the whitelist. The cycle also includes data cleaning to remove spelling and typographic errors. First the whitelist was seeded with an existing list of 5000 common English words. This list was then compared against the police free text and those words that were not on the safe word list were counted and presented in a frequency table. Words in the frequency table are reviewed and those that were not names are added to the safe word list.

8.5.3 Base word list

In order to seed the process of generating the safe word list a base list of common English words is required. There are numerous word lists that have been created, however they tend to have a flaw for this usage, and that is that they have been generated from data, for example Wikipedia, that contains names. What was required was a word list that was not just generated through simple frequency lists, and so was unlikely to have common names.

The Oxford 5000 [1] is a list of what are thought to be the most important words to learn for those learning English, as it is not based on word frequency it did not contain common names, therefore this list was used as the base for the list of safe words. As part of forming the base word list, the Oxford 5000 was compared against name lists, principally name lists from the Office for National statistics [2] that contain all forenames and surnames used in England and Wales, to see if words that could be names were used in the list. Throughout this process, because of the variety of names that can be used, there needs to be a balance of risk. As an example in the ONS list of forenames the name “A” is given, clearly because “A” is such a popular word and a very rare name then most, if not all, usages of the word “A” in the MO text are not likely to be referring to an individual.

Therefore all words from the Oxford 5000 base list were checked against the name lists and a judgement made as to whether the word should remain in the safe list or not. Although 220 of the words were also in the ONS names lists no words were removed from the base list as they were all deemed relatively obscure names.

8.5.4 Developing the safe word list

The safe word list was further developed by comparing the safe word list with all unique words in the police free text see Figure 8.1 with a minimum frequency greater than 10. 10 was chosen primarily due to time available to clean the data. If the word from the police text was not in the safe word list then it was manually reviewed by the author. If the word was deemed sufficiently

[1] https://www.oup.com.cn/test/oxford-3000-and-5000-position-paper.pdf

[2] https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages /livebirths/adhocs/008710babynames1996to2016

safe i.e was not likely to impart personal information then it was added to the safe word list. This allowed more of the free-text through the de-identifcation process.

The unique words generated from the police text will contain normal English words, police jargon and misspellings. Additional words that were added to the safe word list were called *police words* , as they had been generated directly from the police free text data. Examples of *police words* include “complainant”, “stated” and “suspects”. 5205 additional police words were added to the base word list.

This process of adding additional words to the safe word list was only completed with, and therefore tailored to, the MO data. There was insufficient time to tailor the process or the resulting word lists to the police incident data. When the police incident data was de-identified it was completed using the wordlists generated from the MO free text data. The effect of this is to remove more text from the incident logs than is necessary. As an example the “:” symbol was not used in the MO texts and so was not added to the safe word list along with other standard punctuation, whereas it is used in almost every incident log. Therefore every incident log now has the redacted symbol “XXXXX” instead of every “:”.

Once the safe word list was generated it was then used to de-identify the police text. This step is explained below.

8.5.5 De-Identification Process

Once the safe word list had been produced the final data cleaning and de-identification process was completed in the following steps, see Table 8.8 for an example output:

1. Homogenise text. Tidy the text to remove unnecessary pluralisation’s, change jargon and correct common spellings.

2. Replace Information. Use pattern identification to replace known data types with their category so for example replace an actual number plate value with ’NUMBER_PLATE’

Example MO	Suspect(s) unknown steal zebra pattern clothes and hit vctim. They leave in a car vl51pld towards big hill. Hitting Donald Trump as they flee.
De-Identified MO	Suspect unknown steal XXXXX pattern clothes and hit victim. They leave in a car NUMBER_PLATE towards big hill. Hitting XXXXX XXXXX as they flee.

Table 8.8: This table depicts a single example MO, not real, before and after the de-identification process.

3. Whitelist the text. Check every word in the text with the safe list. The safe list is made of the original base word list and the police words. If the word is in the safe list it is allowed to remain. If the word is not on the whitelist then it is replaced with 'XXXXX'

8.5.6 Data Security

Despite best efforts, the possibility remains that a person's name will slip through the net if it is made up of normal words e.g. May Summer. However, an additional procedural control was used to reduce such risk. This measure was the data infrastructure used to secure the data and procedural process. This data infrastructure heavily restricted data export and only allowed access to the data by named members of the research team.

8.5.7 De-identification Results

The de-identification process was not formally tested with the MO data. That is, the MO data was not systematically searched for personal data to determine what percentage of personal data remained after whitelisting. What is known however is how much of the original text data was recovered. This metric is of interest because significant effort is made to develop more sophisticated techniques principally to reduce the result of false-positives i.e. removing non-personal data. For example in Table 8.8 it can be seen that the word zebra is removed because in the context of police MO text "zebra" is a rare word. However, arguably there is no need to remove this word and to do so can unnecessarily lose information on which to fine-tune the PTMs.

For police MO texts the data recovery rate was 97%, that is 97% of the words, by volume, used in the police MO texts were not redacted. The police incident log text was lower at 92%. The police incident data was expected to have a lower recovery rate for two reasons, firstly the word lists were not optimised on the police incident text and secondly the police incident text is expected to contain personal data and so more text is expected to be removed. The MO data retrieval rate was high, and anecdotally from the those researchers that read the texts, comprehension was not overly affected by the redaction of words. The police incident data redaction rate was higher, because the incident text was already noisy and unedited, the impact on comprehension is difficult to judge, but was certainly thought to have had a greater impact on comprehension than the loss of words for the MO data did.

8.6 Data Limitations

The data used in this thesis was limited in a number of key ways. The limitations are generated throughout the data generating process right up to and including the choice of language model used. The key limitations are highlighted below:

1. Police Data Coverage. Police data does not cover all crimes committed and the paucity of coverage is non-random. This non-random coverage is not new and is well documented (Tarling & Morris, 2010a). However it does mean that any patterns or insights drawn through using these techniques with police recorded crime will be subject to these same biases. This is a well known problem and is also a problem when using police structured information.

2. Information completeness. The texts that are provided are not complete representations of the crimes or incidents that they describe. This incompleteness is in some ways deliberate, the police officers or staff only report positively not negatively e.g. they do not generally report on what doesn't happen. Secondly the completeness may be non-deliberate through bias, Officers can only report what they know and as it is widely reported that certain sections of the community to do not engage

as fully with Police Officers as others (Buil-Gil, Moretti, & Langton, 2021). It is entirely plausible that police crime descriptions, and therefore the information that they contain are biased. A future area of study would be to analyse crime descriptions across victim and geographical characteristics to ascertain if they are systematically different in their percentage coverage of the key facts.

3. De-identification. The de-identification process will have removed information from the police texts. this was an unavoidable step for this research in order to provide a reasonable level of data security. The information removed will also have been biased towards rarer words, as the de-identification process was biased to keeping more popular words. Although it is worth noting again that police staff using these models on their own data within their own systems would not have to complete this step.

4. Model compatibility. PTMs have a list of words that they are trained to recognise. If a word is used that is not on that list then it is broken down into word pieces that can then be recognised. As PTMs were not built on police data there are certain words (see Table 8.7) that are not recognised by the language model but are frequently used by the police to convey information. As these words will be broken down into pieces it is plausible that meaning will be lost. The impact of breaking down specific informative words is unknown and is important issue for future research.

All of the factors above will have contributed to limitations within the data. Some of those factors are inherent to police data and have been well studied for examples issues surrounding data coverage, however other issues such as biases in police textual data are not well studied and will require further study to understand the extent of the bias and errors that they may introduce.

8.7 Conclusion

This chapter has introduced the data that is to be used in the resulting studies. All of the data to some extent was changed in order to facilitate research access. As discussed, the changes are likely to have a relatively minor detrimental affect on the ability of the language models. The next chapter will explore the methods that were used across each study with the data presented here.

Chapter 9

Methods

9.1 Introduction

This chapter explains the methods used to extract information from the text data and understand how well the PTMs worked. It will focus on three stages of building and checking the language models. Firstly the approach for labelling the data, active learning, will be explained. Labelled data is required as supervised language models require example texts with the correct label to be presented in order for the model to learn the patterns. Secondly, once the data is labelled the model is fine-tuned on the police data so that it can discriminate between texts and classify them appropriately, this will be the focus of the second part of this chapter. Finally the chapter will address methods to understand how well the models have performed, using the ALGO-CARE framework, introduced earlier, to guide the selection of metrics. Within each study chapter variations from methods explained in this chapter will be stated.

9.2 Data Labelling

In order to use supervised machine learning techniques a portion of the data has to be read and assigned the correct label so that the PTMs can be fine-tuned. Deciding which data to consider for labelling is an important process and a technique to assist with this, called active learning, was introduced in the Chapter 3. Active learning relies on labelling small batches of data then using the PTM in question to find those unlabelled data that it is most uncertain about for the next batch of labelling. This section will explain the labelling and active learning processes used to label the data for the different studies explained in the forthcoming chapters.

9.2.1 Labels

Throughout this thesis language models are going to be used to classify texts, and so the labels that must be given to the data are generally labels that either include or exclude a text from a particular classification. As an example of labelling a burglary MO was read and it was labelled either as having a car stolen or not having a car stolen, thus the classification was "car stolen" and within that classification texts were either labelled "1" if they had a car stolen or "0" if they had not. If it was unclear if an event had happened then it was assumed that it hadn't happened. The same MO text would also be labelled for other events such as the use of force. Each set of classification label e.g. car stolen labels, had to cover all eventualities that could be contained within the text. For the PF1 data only the author labelled the data. For the PF2 data two data labellers were employed, both labellers labelled the same data. Where there were disagreements between the labellers the author adjudicated.

What To Label For?

For both sets of data the following process was generally followed for establishing what subject to label for and how to assign labels. Firstly labels were suggested based on the hypothesis to be investigated or the problem at hand. Once the first suggestion of labels was made a random selection of texts were read. This first pass of the data was to ensure that the text covered the event of interest and secondly to form labels that would cover all eventualities for that event. Once this was completed a practice session was then held with all labellers to run through and discuss a random sample of texts. For the ASB practice sessions, the practice session also had a former Detective Inspector present to assist with the discussion and decisions.

Labelling Implementation

Once the practice sessions were completed the labellers would then label the texts in batches of one hundred texts for MO data and fifty for police incident data. Labelling was completed online and in isolation. Occasionally there was feedback to the labellers if the need for clarification arose as a result of unforeseen results within the text. For both batch types (MO texts of 100 and incident texts of 50) it would generally take around 1 hour to label a single batch. The data labelling was completed asynchronously, which meant that labelers were free to label at a time convenient for them - however this had the effect of elongating the labelling cycle as moving onto the next stage can only be completed when all labelling of each batch was complete. Labelling was conducted in Excel sheets with conditional formatting that only allowed pre-allocated responses to be selected.

The first few sets of texts were randomly chosen for labelling, after which an active learning strategy was then used to choose the most important texts to label from a modelling perspective. The next section explains in more detail how the active learning strategy was used to select the texts for labelling.

9.2.2 Active Learning

Active learning was introduced earlier in Chapter 3 and is a technique to reduce the labelling burden required for training a model by seeking out examples that will help the model improve the most. As labelling texts is a resource intensive procedure, active learning was used to reduce that burden by more intelligently select texts to be labelled to help improve the model accuracy more quickly than by random selection.

General Process.

Figure 11.1 depicts the general process for the active learning strategy. Starting in the top left corner. As a reminder three data sets have to be built for the modelling process. These are:

1. Test set. Picked randomly. Used to estimate the effectiveness of the

trained PTM. This data set is never seen by the model during training and so the data is new to the model at test time.

2. Validation set. Picked randomly. Used to help tune parameters for the PTM. This data is used during the training of the model to gauge progress but it not used directly by the model for fine tuning, but rather to prevent overfitting.

3. Train set. Selected through active learning. Used to train the PTM. This is the actual data that the language model will be fine tuned on and directly influence the models classifications.

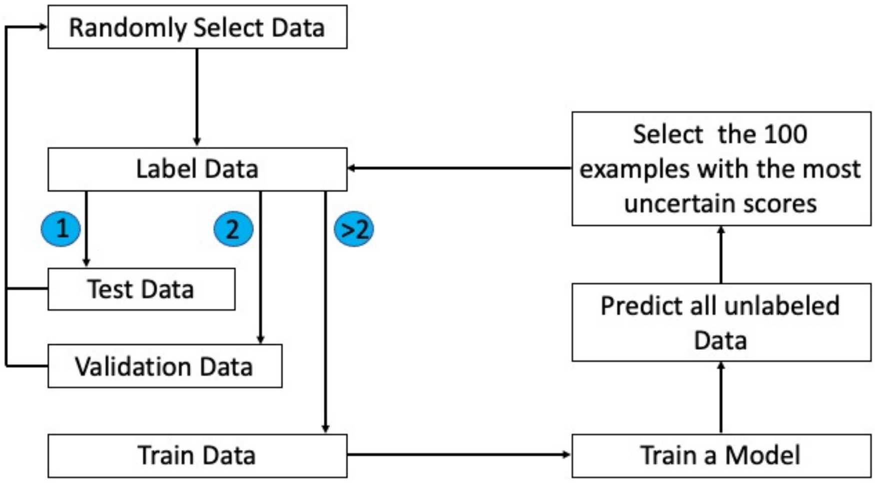


Figure 9.1: Active Learning Process. Start by randomly selecting n data and labelling. Steps 1 and 2 randomly label Test and Validations sets. Step 3 uses randomly labelled data to train a model and to classify all unlabelled data. The texts with the most uncertain scores are then labelled and added to the train set to further fine-tune the model. Data is iteratively added to the data set until the model has satisfactory performance.

The first to steps in the process are preparatory and they are to randomly select data to be labelled for the test and validation sets. The third step is the final random selection, and this random selection selects the first batch of texts for the train set. Once selected these samples are labelled and used to fine-tune a model. The trained model is then used to predict all MO texts that have yet to be labelled. Once completed the results of the model predictions are then used to discover which of the MO texts the model was most uncertain about.

Quantifying model uncertainty was achieved by ordering of differences in output probabilities. PTMs output log-probabilities for each class. The absolute value of the difference in log-probabilities are then ordered and the MO text relating to the smallest values are selected. This process finds the texts that the model is most unsure about. These selected samples are labelled and then the train, predict, select cycle is repeated. This cycle selects the hardest to label texts on each occasion until the decision is made that the model no longer needs to be fine-tuned. Once this decision is made the active learning process stops.

Batch size

The first decision for an active learning strategy is to decided the batch size, how many texts will be labelled in each sitting. Active learning can be completed with a batch size of 1 allowing for a selection after each text has been labelled. however as the labelling in this research was being completed asynchronously and generally by more than one person this would have led to a very slow labelling rate. For this study the active learning was conducted in batches of 100 texts. 100 texts were selected because it translated into a suitable length of time to devote to labelling data - around 1 hour. Much longer and concentration and accuracy may have been degraded, any shorter and the overall labelling rate will have been degraded.

9.3 Pre-trained Language Models

In all three studies in this work the modelling utilised PTMs that were introduced in the first part of this thesis. PTMs have been trained on large volumes of generic texts to give them a broad understanding of language. These language models are then further fine-tuned by exposure to police texts so that they are then able to classify the police texts as required. Each classification type requires a different fine-tuned PTM, so although all of the burglary texts were classified using a PTM, there was a separate model fine-tuned for each classification type or question. So for example there is a PTM fine-tuned for classifying if a motor vehicle was stolen and a separate PTM fine-tuned for if force was used. As previously mentioned PTMs can be useful in the context of the analysis of police free text data, because they can be utilised with little feature engineering effort as they already have a general language understanding, this is in contrast to other general machine learning models that do not have this understanding pre-built and would therefore require extensive feature engineering thereby increasing the time and technical burden on the police analytical staff.

Throughout the studies completed here the modelling was classification modelling, that is take a single piece of text, for example a MO text and classify it as either belonging to a labelled group or not. For example the label could be car stolen and each text would either be classed as having a car stolen or not. The type of language modelling task that is required to be completed influences the selection of PTM. For classification tasks encoder models, of which BERT is the most widely used, are the most appropriate selection as they are able to encode the information from the text into a single output - the classification (Qiu et al., 2020).

PTM were utilised through the Transformers package (Wolf et al., 2019) in python. The Transformers software package allows modern transformer PTMs to be used within a simple interface. The package includes the language models as well as the surrounding architecture to utilise the models such as the tokenisers to prepare the text and interfaces to quantify the performance of the models.

Within this thesis two PTMs were used. Firstly BERT was picked as at the time of commencement of this study it represented the most advanced encoder style PTM of its class and was widely regarded as the most capable PTM (Qiu et al., 2020). One weakness of BERT however is that it cannot handle long texts, so for the police incident texts another PTM had to be utilised that specialised in computing longer texts. For this reason the Longformer PTM was selected for use with the police incident text. The next sections will introduce these models and explain how they were used.

9.3.1 BERT

BERT was first introduced in 2018 (Devlin et al., 2018) and immediately made an impact in the field of NLP by providing new state of the art scores in a set of benchmark NLP standards, known as the GLUE (General Language Understanding Evaluation) tests (Wang et al., 2018).

This section will first describe BERT, how it is built and the inputs it uses and the outputs it generates. Since inception there have been different varieties of BERT (Rogers, Kovaleva, & Rumshisky, 2020), essentially different parameter arrangements, although there are many similarities between the models for ease this chapter will focus on BERT-large, an original BERT model. BERT stands for Bidirectional Encoder Representations from Transformers. Transformers will be discussed briefly below, but the bidirectional element of the name refers to the fact that BERT can understand context from left to right and right to left, meaning that words can influence the meaning of those words before and after them, just as they do for humans.

This next section will describe BERT the model, how it is trained, the inputs required, the outputs produced and finally how it was utilised for this research.

Model Description

BERT is a deep learning model that has been pre-trained to understand the English language (other languages are available). This PTM can then be further fine-tuned on specific tasks across a spread of varied natural language problems. Unlike other machine learning models that have been discussed this model has two stages for its use, pre-training and fine tuning. Both elements use the same model architecture but the first stage, the pre-training, is much more expensive and time consuming than the second which is why it is fixed. The original BERT model was pre-trained on 16 specialist computers for 4 days, with an approximate cost of $7000 (Devlin et al., 2018). For this reason the pre-train phase of a PTM is not a trivial process. Once the pre-training has been completed the PTM is then fine-tuned on representative labelled data from the target NLP task - in this case police text data.

Training

This section explains in more detail the training phases, pre-training and fine-tuning. It then concludes with a review of the inputs and outputs of the model.

Pre-training The pre-training for BERT is conducted in two parts, recall the purpose of the pre-training is to train the model to “understand English”. To train BERT two representative language tasks were used so that the parameters in the model could be adjusted to correctly encode the information from the input text. Both parts are self-supervised, in that they don't require human labelled data, this means that huge amounts of data can be used without the need for costly human intervention to label parts of the data. The data used for both parts of the training was the BooksCorpus (800M words) and English Wikipedia (2,500M words). The two training tasks were:

1. Masked Language Model. 15% of the tokens from a sentence that is input are randomly masked, these masked tokens are then predicted from the remainder of the sentence. One of the strengths of this procedure is that the model can see both the words left and right of the original masked word, so it can predict the word from all of the context contained within the sentence. This is where the B from BERT comes - because the model can use two directions - bidirectional - to understand each word.
2. Next Sentence Prediction. Sentences were paired from the training data. Half of the time the sentences followed on from each other in the training data, for the other half the sentences were paired at random and so were not semantically paired. The model had to predict, given the first sentence, whether the second sentence actually followed the first. This has the benefit of training the model to understand the relationships between sentences as groups of words.

Fine-tuning Fine-tuning is used to adapt BERT to different and specific NLP problems. This can encompass a wide variety of problems such as question-answering or Named Entity Recognition. For our purposes the fine tuning is for classification, and for each classification task a separate instance of BERT was tuned. In order to fine-tune a BERT model for classification an additional classification layer is added as the new final layer in the BERT model. The weights for this final layer are then adjusted as the model learns from the training data presented to the model. Data is presented to the model in batches, a total presentation of all the training data is known as an epoch. There can be multiple epochs in each training cycle. Too many epochs though and the risk is that the model over fits to the training data. A model that has over-fitted to the training data does not generalise well to unseen data. Validation data is used, as a way of detecting the over fitting, and therefore to gauge the number of epochs to use. After fine-tuning the model is ready to be used on unseen instances. The next section explains the inputs of the fine-tuning process and the resulting outputs.

Inputs and Outputs

BERT does not directly take words as inputs it takes tokens after the texts have been through a tokeniser. A tokeniser takes a sentence as an input and breaks that sentence down into words that can then be converted to numerical embedding. Not all words are recognised by BERT, in fact BERT only has a vocabulary of 30522 words (Nayak, Timmapathini, Ponnalagu, & Gopalan Venkoparao, 2020). If a word is not in the BERT vocabulary then the tokeniser will break the word down into recognisable tokens which can in fact be word pieces like “int” or “un” as well as words. This process helps to make BERT robust to previously unseen words. For instance untidy is not in the BERT vocabulary so it is broken into wordpieces “un” and “tidy”. This can be a problem because the summation of the word pieces does not always equate to the semantics of the original word and so meaning can be lost (Nayak et al., 2020). Once the tokenisation has occurred the words are converted into word embeddings that are vectors 768 numbers in length. As mentioned previously these word embeddings have been built to numerically encode the semantic meaning of each word. It is these embeddings that are then fed to the BERT PTM as the inputs.

Once BERT has been trained and fine tuned the area of interest is the output, as this generally contains the information of interest for the task at hand. For each NLP task there is a different model added to the PTM. For classification a classification model is added to the PTM. The model takes the final information from the BERT model (encoded into a single classification token) and uses a linear classifier model to change the output into probabilities for each potential classification. The final classification is selected by picking the classification with the largest probability. The output from the BERT PTM is a probability for each possible classification.

Utilising BERT.

The first step of the utilising the model is to tokenise the input text. This tokenisation takes the input text e.g. MO text and splits the text into tokens. Most of these tokens will be the words, but as mentioned above BERT only recognises 30,522 words and so some words will not be known. These unknown words are split into word pieces that are known. Unlike other NLP models there is no standardisation of the language through shortening words to their lemma, or removing stop (high frequency) words. One choice that is available is either to keep using cased words or transform all letters to lower case. As these documents are typically not edited the tokeniser was set to change all letters to lower case.

Before initiating the BERT model hyperparameters that govern the models behaviour have to be selected. Essentially hyperparameters exist because there is no proven way to optimise how a PTM learns given the data it is to be trained on. As mentioned earlier these hyperparameters include the size of the data batches as the data is fed to the model, the learning rate, how quickly the model changes adjusts to the data it has seen and the amount of times the model sees each piece of data (number of epochs). The batch size was set at 16 the lower of the two recommendations in the original paper (Devlin et al., 2018). For the learning rate we again choose the smaller recommended value (2e-5) the tuning of which is governed by the recommended Adam optimiser. To compensate for the lower learning rate we choose a larger value for the number of epochs than that suggested to ensure the models do not stop training short of a good solution. Initially the number of epochs was 8 but that is reduced on a per model basis as necessary with feedback from the validation data.

For each epoch the validation data was used to compute model metrics. This allowed a view of when the model had stopped showing general improvement and was then overfitting to the training data. Once the training was finished the validation set labels were computed by the model to produce classifications for each text in the validation set, typically 200 MO texts or 100 incident logs. Theses classifications were then used to compute model metrics. As discussed in part 1 the metric selected was the Mathews Correlation Coefficient (MCC) which is robust to imbalanced class problems. As an additional step for the active learning sequence the model that had just been trained would then also label all of the remaining unlabelled data so that the next batch for labelling could be selected.

Models can then be saved to a hard drive much like other files for later reuse if required. Typically only the model weights are saved not the entire model framework.

Once the active learning had finished, and no more data labelling was to be completed, the model was ran ten times on the final training set. BERT models, as with all deep learning models, have random elements to the training process so the results can be slightly different on each training run. Therefore the final training was completed 10 times with the best model being selected by the resulting MCC metrics.

9.3.2 Longformer

BERT models are powerful, but they do not scale well for longer pieces of text, for that reason the BERT model is designed to only take up to 512 tokens as an input. Some researchers have previously circumvented this limit by splitting longer pieces of text into two documents, running the model for the two documents then combining the output, however as context from one part of the document may no longer affect the second this approach is seen as sub-standard (Beltagy, Peters, & Cohan, 2020a). For this reason BERT models were not suitable for the longer police incident log text which as we have seen can be over three times longer than the MO text.

The Longformer model (Beltagy et al., 2020a) was therefore chosen to classify the police incident text as this architecture is designed for longer pieces of text. The Longformer models use a very similar architecture to the BERT models, in that they are both based on the transformer architecture. The Longformer models, however, have modified the method for calculating Attention. Attention is the method for identifying contextual information across the whole text sequence. Calculating Attention in BERT is quadratic to sequence length, but in the Longformer architecture the model architects were able to modify the calculations so that it can now be calculated linearly with sequence length, though with some loss of specificity (Beltagy et al., 2020a). Thus they are able to accept longer sequences of texts. The Longfomer models were used in the same way as the BERT models, with a separate tokeiniser provided by the transformers package for data preparation.

Even though the Longformer architecture is designed for longer pieces of text the models are still computationally expense to run. In order to minimise the computational expense the batch size was reduced to 8 to reduce the amount of texts that the model would consider at anyone time. The remainder of the models hyperparameters remained the same as the BERT model.

All of the PTMs used in this research have now been introduced and an explanation of how they were utilised given. The next section reviews how the models performance was judged after they had been trained on the training data set with hyper parameters selected with the validation data.

9.4 Model Performance

The previous sections have explained how the data was labelled and how the NLP models were trained. This section will now explain how, with a trained model, the performance of that model was explored. Typically performance, especially in computer science, is heavily predicated on how correct the model was. In essence consideration is only given to accuracy and similar metrics such as MCC. Here though we recognise that for a model to be used in service with the police, and most likely all public service settings, the model needs to be more than just accurate. Using the ALGO-CARE mnemonic introduced earlier, we see that accuracy is only one factor in a list of eight factors that are described for using algorthims in a police context. As before we highlight the two additional factors described in ALGO-CARE that pertain to the implementation of these models. Firstly the framework asks "Is appropriate information available about the decision-making rule(s) and the impact that each factor has on the final score or outcome?" in relation to how explainable the results are and secondly "What are the post-implementation oversight and audit mechanisms e.g. to identify any bias?" as the results should be challengeable. With these factors in mind the performance of the models will be further explored through explainability and bias. These investiagtions are discussed in more detail below.

Throughout the research model performance will be judged on a randomly selected test set. Test sets were randomly selected from the data before any data was removed for training. None of the test set will have been used for either the training or the validation of the model fine-tuning. This means that during performance assessments the test set is new to the model. Therefore, metrics from the test set provide a reasonable assessment of how the model will perform on unseen texts.

$$MCC = \frac{(TP * TN – FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{9.1}$$

Where: TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

9.4.1 Metrics

As mentioned in the first part of this thesis, there are a multitude of different model metrics that can be used to judge a models performance. Selection of the model metrics should be based on the type of problem and dataset used. In this instance the problem was one of classification with an imbalanced data set i.e. one of the potential classifications was much rarer than the other. As outlined earlier Mathews Correlation Coefficient (MCC) is a good metric for this type of problem as it gives a standard score between 0 and 1 independent of the number of classification categories i.e binary or across more than 2 possibilities. Secondly unlike more basic metrics such as Accuracy the metric is able to account for imbalanced classes where rare instances may be difficult to predict. MCC was calculated using the scikit-learn package in python (Pedregosa et al., 2011), the equation for MCC is given in Equation 9.1. This metric is the primary metric for understanding how well the PTM got the classifications right overall. However, also of interest is how did the PTM came about its classifications (explainability) and how well did the PTM do across groups of instances within the dataset (bias).

9.4.2 Explainability

As explained earlier in Chapter 4 how a model came about its predictions is just as important as if it got the predictions correct, as being able to explain how a model is making decisions builds trust in the model. To be explainable the model must be able to explain or show why certain classifications were given. As mentioned in Chapter 4 these can either be global explanations, where the model can be explained for every possible input, or local explanations where the explanation is centred on the individual instances to be classified. For deep learning and in particular NLP it is very difficult to produce global explanations because of the vast array of possible inputs, for this reason we focus on local explanations using the LIME package introduced earlier.

LIME

LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) is an algorithm that is used to explain why a model has made a certain prediction. In essence the model takes a real individual instance to be predicted then modifies that instance slightly, in the case of texts it removes one or more words. The model of interest is then re-ran on the modified instance and the new output noted. Recall that in classifications models the output is a set of probabilities for all classifications and not just a single classification. So even if the final classification has not changed, it is likely that the underlying probability of that classification will have changed. Modifications (of the same instance but modified in a different way) are repeatedly selected on a number of occasions so that a local representation of a number of similar but distinct instances can be built. With these modified instances and their resulting probabilities a simpler local model, such as a linear regression, can be built that is then more easily interpretable. The coefficients of the resulting regression can be used to understand the effects of the modifications and therefore of the feature modified on the final probability. Thus at a local level the prediction can be explained by how much a particular feature (or word) is responsible for changing the probability. See Figure 9.2 for a simplified pictorial example, where the bold red cross is a whole MO text, and the smaller red crosses would be the text with some of the words removed. The black dashed line is then the linear model from which the coefficients of the removed words can be deduced and their impacts understood.

LIME Implementation.

In order to get a view of explainability the LIME model was ran on all of the test set after the final classification model for each problem. For each MO or incident text random perturbations were conducted 100 times (selected based on trials,

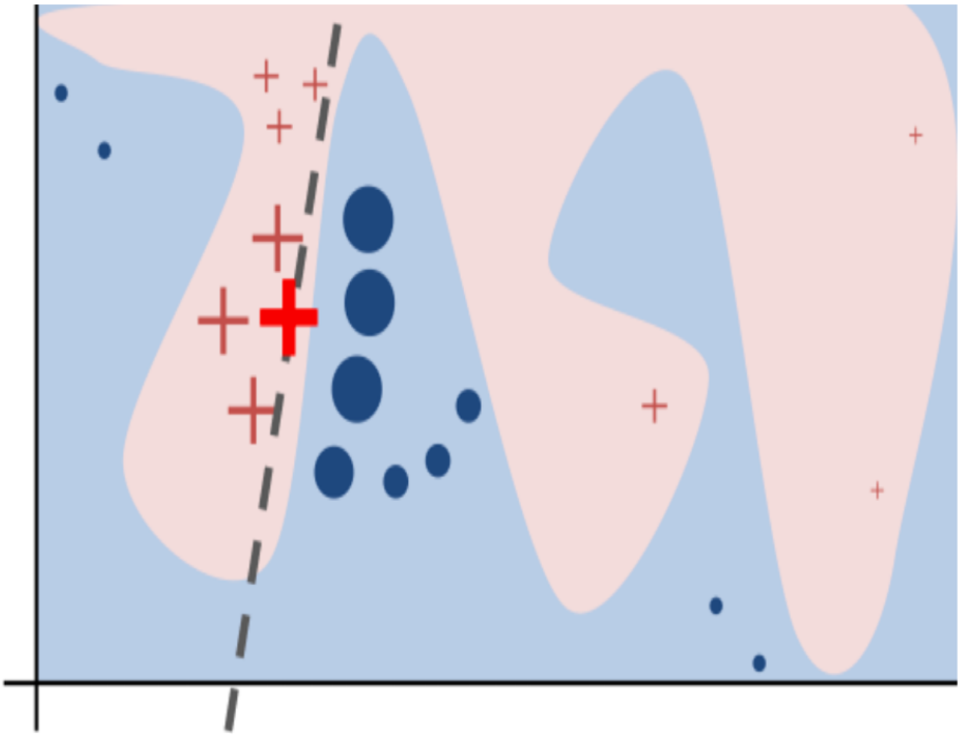


Figure 9.2: Toy LIME example. The bold red cross is the original unmodified instance (original text) , smaller red crosses are the modified instances (texts with a word removed). The resulting black line is from the regression and is the learned explanation that is locally faithful i.e built on a single text. The true complex decision boundary is represented by the pink/blue background and is true globally, although generally unknown. Reproduced from (Ribeiro, Singh, & Guestrin, 2016)

there was little variation in output at 100). This 100 iterations produced a single linear model for each MO text. Once complete the coefficients from each of the resulting linear models were pooled across the entire test set so that a broader view could be taken on the words that were most important for the classifications. This data is then presented in a word cloud, where the size of the word is related to the how important that word is for the final classification in the whole of the test set. The larger the word in the visualisation the more important it was for classification of the police text in that problem. If the words make sense to a human for the classification, then it is likely that the model is using the words to form a judgement in a similar manner to how a human would use them. However, if the larger words don't seem sensible for a classification then it maybe that the model has picked up on a spurious correlation in the training set. These visualisations then allow a judgement to be formed on how the model is working - if this is is inline with human expectations then the model can be considered more trustworthy than had it not been.

9.4.3 Bias

In Chapter 4, three broad areas were identified as sources of bias. i)Data coverage relating to the inconsistencies of reporting crime to the police. (ii) Data completeness - where the police may or may not systematically record different levels of detail about certain crimes or from certain sections of the community. (iii) Finally algorithmic bias was introduced, which given the data, was the algorithm making more or less errors in certain parts of the data.

The first two are difficult to quantify in the study because the only data available is the police data. We do not have access to the totality of crimes conducted, nor do we have access to perfect descriptions of the crime to understand if there are important elements systematically missing from the police descriptions. The final type of bias, algorithmic bias is within the gift of this research to identify and is an important element for consideration.

Algorithmic bias typically occurs in PTMs because of the data that was used in the pre-training phase and how that relates to the data being analysed. For instance if certain police texts are very dissimilar to the pre-training data then they may not be classified well, additionally if there is bias within the pre-training data, that may be carried through to classifications biases of police text data. For the purposes of this research we split this bias into two categories those relating to the text and those relating to the crime or incident being described.

Firstly there are qualities of the texts themselves that can be described through statistical variables - statistics that are produced from the texts that the PTM is used with. Examples are the length of the text and the amount of out-of-vocabulary words. Secondly there are characteristic variables of the crime being described that the model may or may not be able to deduce from the text that is used to train the model. For instance the location of the crime or the gender of the victim. Of course these two types of variables may not be independent of each other, for instance it is possible, though not evidenced, that certain victims groups may have shorter crime descriptions because of the relationship they have with the police. These two variable types are used to explore potential biases with using PTMs.

In the literature bias is often measured through metrics such as extrinsic bias (Goldfarb-Tarrant, Marchant, Sánchez, Pandya, & Lopez, 2020). Extrinsic bias will be used to explore characteristic variables, these metrics are introduced and explored below.

Extrinsic Bias

“Extrinsic bias metrics measure bias in applications, via some variant of performance disparity, or performance gap between groups.” (Goldfarb-Tarrant et al., 2020). As an example of potential extrinsic bias from this research, if a burglary classifier had higher error rates for male victims than female victims then this would be an example of an extrinsic bias. Goldfarb-Tarrant et al. (2020) identifies the two most popular metrics for investigating extrinsic bias, these are listed and explained below.

But before getting to those definitions we need to remind ourselves of two more basic definitions, *Recall* and *Precision. Recall* can take a value between 0 and 1 and represents the percentage of positive instances that have been returned by the model. *Precision* can also take a value between 0 and 1 but in this instance it represents the percentage that were actually positive from those identified as positive by the model. Now the two extrinsic bias metrics are explored.

$$P(\hat{Y} = 1|A = x, Y = 1) = P(\hat{Y} = 1|A = y, Y = 1) \tag{9.2}$$

$$Recall_x - Recall_y \tag{9.3}$$

$$P(\hat{Y} = 1|A = x, Y = 0) = P(\hat{Y} = 1|A = y, Y = 0) \tag{9.4}$$

$$Precision_x - Precision_y \tag{9.5}$$

- **Equality of opportunity.** Equality of opportunity occurs when Equation 9.2 is satisfied (Goldfarb-Tarrant et al., 2020). That is where the probability of being classified as positive ($\hat{Y} = 1$), given that the sample is positive (Y = 1), is the same regardless of what group the sample is drawn from (A = x or A = y). Equation 9.2 is based upon recall and therefore Equality of Opportunity can be measured through Equation 9.3. Where $Recall_x$ represents the recall from the reference group (sometimes considered the privileged group) and $Recall_y$ represents the group of interest (sometimes referred to as the underprivileged group.(Hardt, Price, & Srebro, 2016)

- **Predictive parity** (Verma & Rubin, 2018). Predictive parity is similar to equality of opportunity above, but relates to the probability of incorrect predictions as seen in Equation 9.4. In this case parity occurs when the precision from each group is the same, that is the probability of being identified, given that it wasn't a positive instance is the same regardless of the group the sample is drawn from. Again a simplified form to calculate the metric is given at Equation 9.5

  These extrinsic bias metrics were calculated for each test set. However the weakness with this approach is that only a single data point is obtained on which to judge bias. In order to provide more evidence, rather than just a single metric, a cross-validation process was implemented to provide a bias estimate with confidence interval. This cross-validation process is explained next.

  For the cross validation process 20% of the available labelled data was randomly selected (available includes all data labelled for the test, validation and train data sets). This 20% was used as the test set. The remaining 80% was used as the train set. There was no validation set as the hyper parameters were fixed. A PTM was fine-tuned using the 80% train set then used to label the 20% test set. Bias metrics EoO and PP were then calculated on the 20% test set. This whole process was repeated 10 times so that there were 10 sets of bias metrics.

For each bias metric a non-parametric hypothesis test of equal means was conducted for each metric, testing if the mean of the metric was 0 or not using all ten data points. A significant p value would indicate bias at a statistically significant level across the experiment. The mean of the ten metrics indicates the direction and the size of the bias.

9.5 Summary

In this chapter the main elements of the method have been set out. These methods will be used in each study and form the basis for the analytical approach. In summary the main steps are:

1. Label the data. The data will be labelled through an active learning strategy. The labelled data will then be used to fine-tune and test the language model.
2. Fine-tune a PTM. The data labelled will be used to fine-tune a PTM, either BERT or Longfomer. This approach has proven to be quicker than building a NLP model from first principles which entails feature engineering.
3. Test. The language model will be tested for performance (using MCC) , explainability (using LIME) and bias (using extrinsic bias metrics) to investigate whether the performance of the models is sufficient for utilisation in a police and POP context.

The next chapters will now introduce each study in turn. Within each study will be a problem introduction, a review of the methods, the results then a discussion of what the results mean. There will be four chapters covering the studies. These chapters will be broken down as follows:

1. Study 1a - Burglary MO data (data - PF1)

2. Study 1b - Active learning
3. Study 1c - Replication study - Burglary MO data (data - PF2)
4. Study 2 - Police Incident texts (data - PF2)

After the studies the final part will be a broader discussion of the results from the PTMs and how or if they might be implemented to assist with POP.

Chapter 10

Study 1a: PF1 Burglary MO

10.1 Introduction

The preceding chapters laid the groundwork for this study. Firstly, the rationale of the study was set out – it is intended to enable POP by identifying intra-crime variation through the use of free-text data. Thereafter, general theories of machine learning and NLP were laid out so as to preface the methods chapter and the general introduction to the data that are employed in the study. This chapter presents the first study of the thesis, and it focuses on the classification of burglary MO texts from the PF1 data. The study sets out to classify burglary MOs by reference to three factors. The first is car-key burglaries, the second is burglaries in which force is used, and the third is burglaries in which only an outbuilding is targeted.

The last category was not examined because there were no outbuilding burglaries in the PF1 data. However, it is explored in the replication study (Study 1c). The next two chapters are also related to the present study. They cover 1) the effectiveness of the active learning strategy that is employed here (Study 1b) and 2) the replication of this study with data from a different police force (Study 1c).

The primary focus of this chapter is on the utility of PTMs for MO data. Utility is examined in the context of three questions. Firstly, can PTMs produce accurate results with police MO data? Secondly, can these results be achieved within a reasonable resource envelope? Thirdly, are the models acceptable for use, that is, are they explainable or affected by bias?

This chapter begins with an overview of the problems and the process by which they were selected. The use of data is explained, and methods are reviewed briefly. One method is added to those that were introduced in the methods chapter, namely the keyword method. This addition facilitates a comparison between the PTM and current police practice. Finally, the results are presented and discussed.

10.1.1 Problem overview

In order to test the utility of the PTMs a selection of representative problems had to be selected. Previous work in the field has focused on burglary data (Birks et al., 2020; Sheard, 2020) in order to explore effectiveness of NLP. This study follows that approach and again utilises burglary data to explore NLP effectiveness. The following sections set out the three classification problems that were identified as suitable for this study.

1. The first classification problem follows on from the work of (Sheard, 2020). In her thesis Sheard states that "this thesis presents empirical evidence that failure to disaggregate beyond official crime classifications risks neglecting heterogeneity of offence characteristics within these. A potential implication of this is that the spatio-temporal parameters on which some prevailing crime modelling techniques are based might not apply to all offences, meaning that any related decision-making could be misinformed". Sheard investigated this matter by showing that car-key burglaries have a different spatial-temporal pattern from non-car-key burglaries. However, in the process, Sheard observed that differentiating between the two types of burglary is laborious and time consuming. She needed much time to complete the process because there were no encoded variables (flags) in the police data that would enable the crimes to be differentiated. More formally, this classification task entailed highlighting burglaries in which a motor vehicle had been stolen. The category of motor vehicles includes cars, vans, and motorbikes. Although the crimes in question are often called "car-key burglaries", if only the keys are stolen. In this thesis the focus was on the the theft of the motor vehicle

as the defining characteristic, not solely the keys.

2. The second classification task originated from a discussion between Dr Daniel Birks and a Detective Chief Inspector (DCI) from Durham Constabulary. Burglaries in an area of Durham had spiked suddenly. A preliminary review of the problem led the DCI to believe that the spike could be attributed to an increase in the incidence of outbuilding burglaries. However, there were insufficient resources to test this hypothesis comprehensively because it would have been necessary to read the available textual information on all burglaries in order to establish a baseline and to identify a recent trend. This problem is representative of the tasks that a police force may wish to undertake. Burglaries of entire homes may be considered more harmful than burglaries that only target outbuildings. Consequently, police forces may wish to understand intra-crime variation and to allocate resources accordingly. The classification task, therefore, was to highlight burglaries in which only an outbuilding (and not a home) had been broken into; burglaries that had targeted both an outbuilding *and* a home had to be excluded.

3. The final classification task was designed to complement the first two. Both outbuilding burglaries and car-key burglaries are relatively rare (approximately 9% of burglaries in the PF1 data involved theft of a motor vehicle), leading to imbalanced classes within the data. A complementary classification with more balanced classes is needed, that is, one in which the act is mentioned much more frequently than the theft of a motor vehicle. A review of a selection of MO texts resulted in the use of force to enter a building being chosen because almost every text appears to contain such information. Use of force is present in approximately 60% of the MO texts, yielding a much more balanced problem. This classification task, therefore, focuses on the method of entry into a building and the use of force. The use of force within the home, for example to destroy furniture, is not included.

The three classification tasks are intended to yield a superior understanding of intra-crime variation within the burglary MO data. These problems are representative of the tasks that an analyst or a police officer may wish to undertake. The next section summarises the data that are used to explore these classification tasks.

MO 1 “Attacked property is mid town house with driveway to the front along with gardens to both front and rear located within a residential area. At time stated person/s unknown go to front door and open letter box and using unknown instrument hook door key from a shelf in the porch. Use same keys to open front door and gain entry remove two sets of car keys from the porch area. Go to a XXXX parked on the drive gain access using keys. Make off at speed with both vehicles direction of travel towsrds XXXX having been disturbed by the occupant.”
MO 2 “Attacked property is a large detached dwelling on a busy road. Property is surrounded by large fences, gates and bushes. Between times stated suspect approach rear patio doors at locus and attempt to gain entry by using mole grip type implement to snap lock. Lock snapped however unable to gain entry. Suspects then use molegrip type implement to snap lock on front porch door. Lock snapped, door opened and house alarm sounds. Suspects jump over wall at front of dwelling, get into vehicle parked opposite and make off down XXXXX in direction of XXXXX.”

Table 10.1: An example of MO texts from PF1 Burglary data. Reproduced from Birks, Coleman, and Jackson (2020).

10.2 Data

The data were introduced in Chapter 8. For this study only the burglary data from the PF1 data was used. The text data was taken from the *Crimenotes* column. Examples of the MO texts can be found in Table 10.1. The PF1 data had already been processed, exact mechanisms unknown, to replace identifiable information with ’XXXXX’. No other data processing was undertaken. There were 9,961 burglary MO texts in total, spanning two years of data.

The data was split into three sets as previously mentioned. The test set was 200 randomly selected texts. The validation set was 200 randomly selected texts. The training set was selected using an active learning strategy. Active learning is explained in the methods chapter and the effectiveness explored separately in the next chapter. In total 1200 MO texts were labelled manually for the test sets.

10.3 Methods

The methods of all of the studies were introduced in the methods chapter. Accordingly, this section only presents a brief outline and highlights points of divergence the stated method.

10.3.1 Labelling

The data were labelled by a single individual, the author. Data selection for the labelling of the test set was completed by using an active learning strategy. The batch size for active learning was 100. Each model (“use-of-force”, “motor vehicle stolen”, and “outbuilding”) was labelled according to a separate active learning strategy. However, in order to generate more labelled instances and because the additional time cost was marginal, every MO that the author read was labelled for all catergories, independently of the applicable active learning strategy. For example, when running the active learning model for the use of force, the MO texts were selected by finetuning the force model. However, each MO text was labelled as “force”, “outbuilding”, or “motor vehicle stolen” when the author read it.

After several hundred MO texts had been read, it became clear that burglaries of outbuildings were not mentioned in any of the MO texts. For this reason, the analysis of that category was discontinued, and no models were finetuned. In total, 900 MO texts were labelled for the motor-vehicle model by using the active learning process, and 700 MO texts were labelled for the use-of-force model. In total, 1,500 burglary MOs were labelled (1500 = 900 + 700 - 100 (because both sets were based on same initial random selection)). The active learning process, and therefore data labelling, were terminated when the MCC for the validation set exceeded 0.9.

10.3.2 Fine-tuning the PTM

Modelling was completed by using the BERT-large-uncased model. The model was utilised through the transformers package on Python, as explained in the methods section. All hyperparameters were set in the manner that was described in the methods Chapter. The hyperparameters were not tuned, except for the selection of epochs (5) for the final fine-tune. A total of 200 MO texts were used for the test set, and another 200 were used for the validation set.

10.3.3 Performance

As described in the methods chapter, three aspects of performance are examined here. MCC is used for overall performance, LIME is used for explainability, and extrinsic metrics are employed to examine bias. Since no victim data are available, the statistical properties of the MO are used to determine whether its length and style of composition have implications for model classification.

In addition to these performance metrics, a fourth aspect was added for the purposes of this particular study. That addition entails comparing the PTM to the workflow that analysts use at present. To that end, a basic keyword search was conducted. The search could be completed relatively easily on readily accessible software such as Microsoft Excel. This keyword modelling process is explained below.

Keyword Model.

The keyword model is based on simple searches for keywords that are likely to feature in the MOs. The model is designed to represent methods that police analysts and/or police officers use at present. Accordingly, it is designed to be relatively simple. The keyword model does not use complex rules in which words can be chained or their presence negated to develop more intricate searches. In essence, if an MO text contains a keyword, the model labels it as a positive example.

The keyword model was developed after reading a substantial number of MOs. Therefore the keyword search may be relatively good compared to one that would have been produced without that experience. The model reflects the manner in which a police analyst may approach their tasks - and so it should be assumed that they will have had some previous exposure to MO texts. The keyword list was built from words that were associated with burglaries where a motor vehicle had been stolen e.g. (car, motorbike). The keyword list was also made more robust to unseen data by adding in a list of popular car brands[1], as the vehicle can often be referred to by brand name alone. The final list of keywords for the motor vehicle model can be found in Table 10.2, A similar process was used for the force model, and the final keywords for that model can be found in Table 10.3. The model was built in R, and it searches each MO description for all of the words on a list. If any of the keywords are present, the MO is labelled as being in the positive class for the corresponding classification model.

Although it was built in R, the model could easily have been created in Excel or through the use of SQL queries (SQL is a database manipulation language with which police analysts are often familiar). In order to fully explore the differences between the keyword model and the PTMs, it was important to track another metric, which is called "recall". Recall was introduced in Chapter 4 and is also explained below. Time was also used as a proxy for effort in order to understand how the burden that analysts must shoulder varies between PTMs and traditional keyword models.

$$Recall = TP/(TP + FN) \tag{10.1}$$

Where TP = True Positive and FN = False negative.

Recall.

Recall is graded from 0 to 1, and it is related to the proportion of examples that are labelled as positives 10.1. A score of 1 means that all of the examples are positive. Recall was selected because it is assumed that the police analysts are interested in finding all ipositive outcomes of a given theme. The downside of using recall alone is that a trivial strategy of labelling all cases as positive would yield a recall value of 1 but would not reflect progress in the eyes of the analysts, who would still be faced with the original sample and all of the negative instances. However, to the present ends, the metric is adequate for exploring the differences between the keyword model and the PTMs.

[1] https://yougov.co.uk/ratings/transport/fame/car-brands/all

Time

Analysts are busy, and the models that they use must not be too time intensive, relative to the results that they offer. Time is used here to understand, at a relatively high level, how much effort an analyst must expend in order to use a model. It is assumed that test and validation sets are required for each model. The training sets need only be utilised by the ML models. During the labelling process, it took the author approximately one hour to read 100 examples. When time is used as a metric, it is important to distinguish between elapsed time and user time. For instance, reading an MO text makes demands on the schedule of the analyst, but waiting for a model to run does not – the analyst can perform another task. Thus, labelling time (user time) is not the same as model training time (elapsed time).

Bias

This section explains how the PTMs were explored for bias. The PF1 data were not accompanied by victim characteristics. Therefore, the investigation of bias in this data set is limited. The PF2 data did include victim characteristics. Accordingly, bias is explored more thoroughly in the following chapters. However, it is useful to inquire whether any of the characteristics of the text data influence the ability of the model to arrive at the correct classification systematically. Three properties were investigated for bias in this study. The first was the length of the MO text. Longer MO texts may contain more information and thus be easier to classify. Recall that BERT can only recognise certain words. Words that are not recognised in their entirety are broken down into pieces until the process can be ran successfully. For example, untidy becomes "un, ti and dy". Pieces of words are investigated by reference to count per MO and as a proportion of the length of the MO text.

The metric of interest is the Pearson correlation coefficient, which allows for the identification of correlations between a statistical property (e.g., length of text) and the accuracy of the classification of each MO text within the test set. Accuracy of classification was defined by reference to the accuracy of the model probability of classification. For example, if the model predicts that an MO text is a positive example with probability of 0.7 and the text is in fact a positive example, then the error is 0.3. Conversely, if the MO text in question

car	dacia	lamborghini	nissan	toyota
alfa romeo	ferrari	land rover	opel	van
aston martin	fiat	lexus	peugeot	vauxhall
astra	focus	lotus	porsche	vehicle
audi	ford	maserati	renault	vehicles
audi	general motors	mazda	rolls-royce	vespa
bentley	honda	mercedes	saab	volkswagen
bmw	hummer	mg	seat	volvo
bugatti	hyundai	mini	skoda	vw
cadillac	isuzu	mitsubishi	smart	
chevrolet	jaguar	motor	subaru	
chrysler	jeep	motorbike	suzuki	
cireon	kia	motorcycle	tesla	

Table 10.2: List of keywords used to populate the keyword model for motor vehicle stolen.

is, in fact, a negative example, then the error is 0.7.

As noted in the methods chapter, in order to arrive at a robust estimate of bias rather than a single value of a metric, a multiple random selection approach was employed to generate a spread – 20% of the labelled data were randomly selected into the test set. The remaining 80% were used to train the model. Once the model had been trained, the 20% test set was used to generate the required metrics. This process was repeated 10 times with different random selections of the test and training sets. Each selection would produce a different value for the metric. Therefore, 10 values were produced for each metric over the course of the experiment.

smash	prized	jemm	forc
kick	break	attack	damage
broken	snap	removed	shattered

Table 10.3: A list of keywords used to populate the keyword model for force used.

10.4 Results

This section will discuss the findings from experimentation of the BERT model with the PF1 burglary data. The section presents findings that pertain to the three main areas of analysis, namely performance, explainability, and bias. The impact of the active learning strategy is discussed in the next chapter. After the findings have been described, their interpretation is presented in the discussion section.

10.4.1 MCC

Table 10.4 displays the results for the force model and the motor-vehicle model. The outbuilding model was not used because of the lack of suitable labels. These results were generated by using the data from the active learning process for the corresponding models. The table includes performance metrics from the PTM and the keyword model. Since separate active learning processes were used for each model, there was an opportunity to combine the data from each active learning strategy to finetune a model on a larger set of inputs. The results from the use of all labelled data (1,500 MO texts) are displayed in Table 10.5. That table has 10 entries because the model was built 10 times in order to obtain an accurate spread of results. Recall that the models were built with an element of randomness; they can be different on each occasion.

Motor Vehicle Model. The BERT model has an MCC of 0.97 and a recall of 0.94 when using only the active learning data (900 texts) are used(Table 10.4). This increases to an MCC of 0.97 and recall of 1.0 when all labelled data are used (1500 texts) (Table 10.5.). In comparison the keyword model achieves an MCC of 0.81 and a recall of 1.0. It is worth noting that the keyword test set

	Motor vehicle		Force	
	Recall	MCC	Recall	MCC
Keyword (Validation)	1	0.65	1	0.56
Keyword (Test)	1	0.81	0.96	0.51
Keyword (Train)	0.97	0.62	0.94	0.45
BERT (Validation)	0.94	0.85	0.94	0.89
BERT (Test)	0.94	0.97	0.94	0.86
BERT (Train)	NA	NA	NA	NA

Table 10.4: Selected metrics from the results of Study 1a. These results are generated only from the MO text that was labeled within the active learning strategy for that model.

result was unusually satisfactory - the validation set and the train set which were used to create the model actually had lower MCC scores than the test set. This is the reverse of what one would expect because the model is usually always better at classifying the data on which it was built.

Force Model. The results for the use-of-force model are similar. BERT has an MMC of 0.86 for the model that is built only on the active learning data (700 texts). When all data (1,500 texts) are utilised, MCC increases to an average of 0.91. The keyword model, when applied to the use-of-force model, yields an MCC of 0.51, which is significantly inferior than that of the PTM. The recall values from the two models, however, are closer to each other. The keyword model has a recall of 0.96, and BERT has a recall of 0.94.

10.4.2 Time.

This subsection compares the time that it takes to build and run each model successfully. In particular, a comparison is made between the PTM and the keyword models. In each case, it is assumed that the test data and the validation data need to be labelled for the purposes of model development.

Therefore, they are not be included in the comparisons. It is assumed that labelling 100 MO texts takes an hour. Additional modelling time within the active learning process is accounted for, as discussed in the next chapter.

Motor Vehicle Model The motor vehicle PTM uses 900 texts. Labelling thus took 9 hours. Fine-tuning the model took an additional 7 hours, although this activity does not call for any human input once initiated – it can be completed overnight or while an analyst is engaged in other tasks. Only 100 MO texts had to be labelled for the keyword PTM. The knowledge that was gained from reading the test set, the validation set, and the initial training set was sufficient to produce a suitable keyword recall model from the validation set. The keyword model therefore required an hour of labelling and an hour of research (to expand the list of keywords so as to include plausible alternatives). Therefore, the keyword model can be built and implemented much more rapidly – user time is 2 hours. Conversely, PTM demands 9 hours of user time and 7 hours of training.

Force Model. Similarly, the PTMs for the use-of-force model took much longer to build. In this case, 7 hours of labelling were followed by 6 hours of finetuning. The keyword model was complete after 90 minutes.

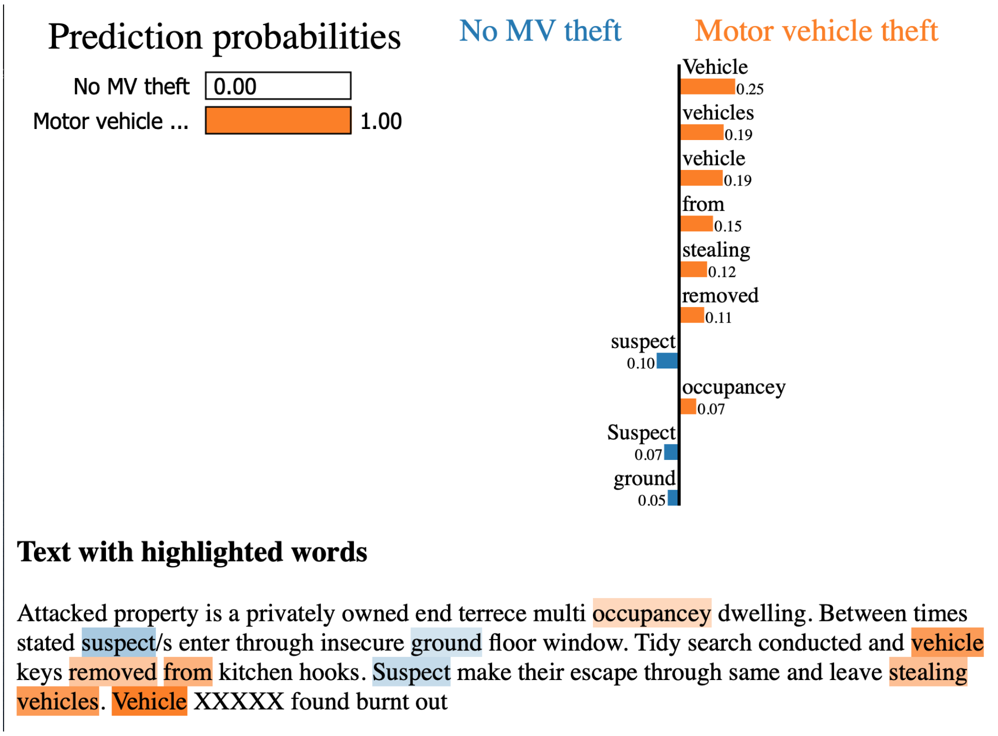
10.4.3 Explainability

LIME was used to examine the PTMs in order to understand which words had the strongest effect on the classifications. Figure 10.1 is an example of the LIME output from a single MO text; only the 10 most influential words are highlighted. The prediction is for a burglary with theft of a motor vehicle. The words that are highlighted in orange contribute the most to the corresponding classification; the words that are highlighted in blue count against it. In this case, the three most important words for the decision were all “Vehicle”.

Although the LIME output that is displayed in Figure 10.1 provides an adequate visual representation of the operation of the model with a single MO text, that style of visualisation does not scale well to multiple texts. Accordingly, a different approach was adopted in order to arrive at a more general representation of the LIME output from several texts. The general

	Force Model		Motor vehicle model	
Run	MCC	Recall	MCC	Recall
1	0.97	0.99	0.97	1.0
2	0.92	0.96	1.0	1.0
3	0.92	0.96	0.97	1.0
4	0.90	0.94	0.97	1.0
5	0.90	0.94	0.97	1.0
6	0.90	0.94	0.97	1.0
7	0.88	0.93	0.97	1.0
8	0.89	0.94	0.97	1.0
9	0.90	0.94	0.97	1.0
10	0.90	0.94	0.97	1.0
Mean(CI)	0.908 (0.89-0.92)	0.948 (0.94-0.96)	0.973 (0.97-0.98)	1.0 (1.0-1.0)
Best Run	0.97	0.99	1.0	1.0

Table 10.5: Each run represents the fine-tuning of a single model using all the labelled data. Each run is independent. Results are different between runs as there are random aspects to fine-tuning that can alter the end result.



Attacked property is a privately owned end terrece multi occupancey dwelling. Between times stated suspect/s enter through insecure ground floor window. Tidy search conducted and vehicle keys removed from kitchen hooks. Suspect make their escape through same and leave stealing vehicles. Vehicle XXXXX found burnt out

Figure 10.1: Lime Output for a single MO text for the motor vehicle theft during a burglary model. The model correctly predicts that a vehicle was stolen. Words highlighted with orange contributed to the positive prediction.

	motor vehicle	Force
MO Length	0.092 (0.150 to 0.009)	- 0.01 (0.071 to -0.099)
Number of Word pieces	0.007 (0.060 to -0.085)	0.001 (0.065 to -0.067)
Ratio MO Length to Word pieces	0.066 (0.141 to -0.042)	-0.004 (0.089 to -0.069)

Table 10.6: This table gives the mean Pearson correlation coefficients between the probability of classification from the NLP model and the metrics listed in the first column. The value in the table is the mean of the ten Pearson coefficients. Figures in bracket are the range.

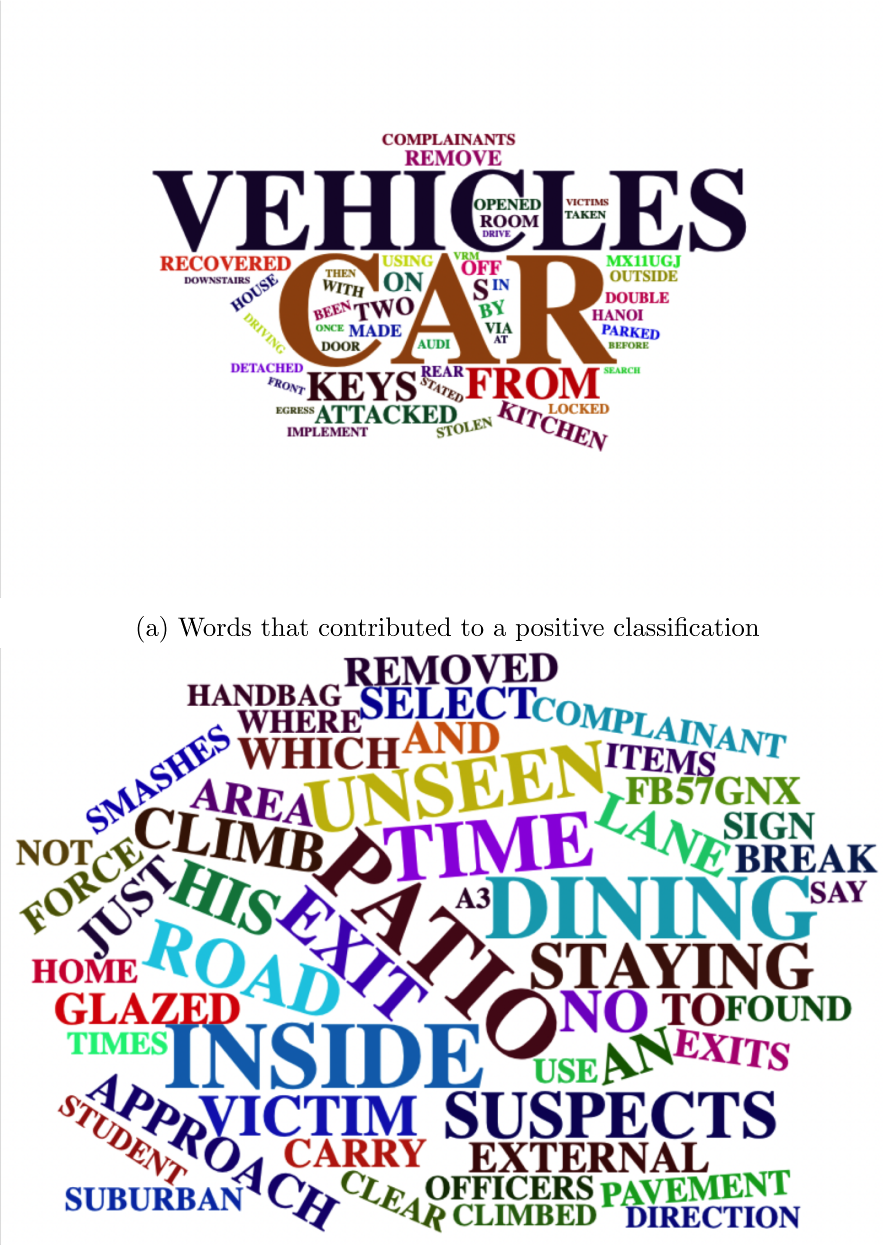
approach was to run the LIME algorithm for every MO text in the test set. The coefficients from the local models that were generated for each MO text were stored, and the word clouds in Figure 10.2 and Figure 10.3 were generated. The size of a word in the word cloud reflects how important it is for the classification of all MO texts in the test set. Word sizes cannot be compared across word clouds.

10.4.4 Bias

Table 10.6 highlights the mean of the Pearson correlation coefficients for the metric in the first column. The mean was calculated from 10 randomly initiated model builds, as described in Section 10.3.3. It is clear from the table that there are no linear correlations between the accuracy of the classifications and the statistical properties of the MO texts. All correlations are very close to zero, as are their ranges.

10.5 Discussion

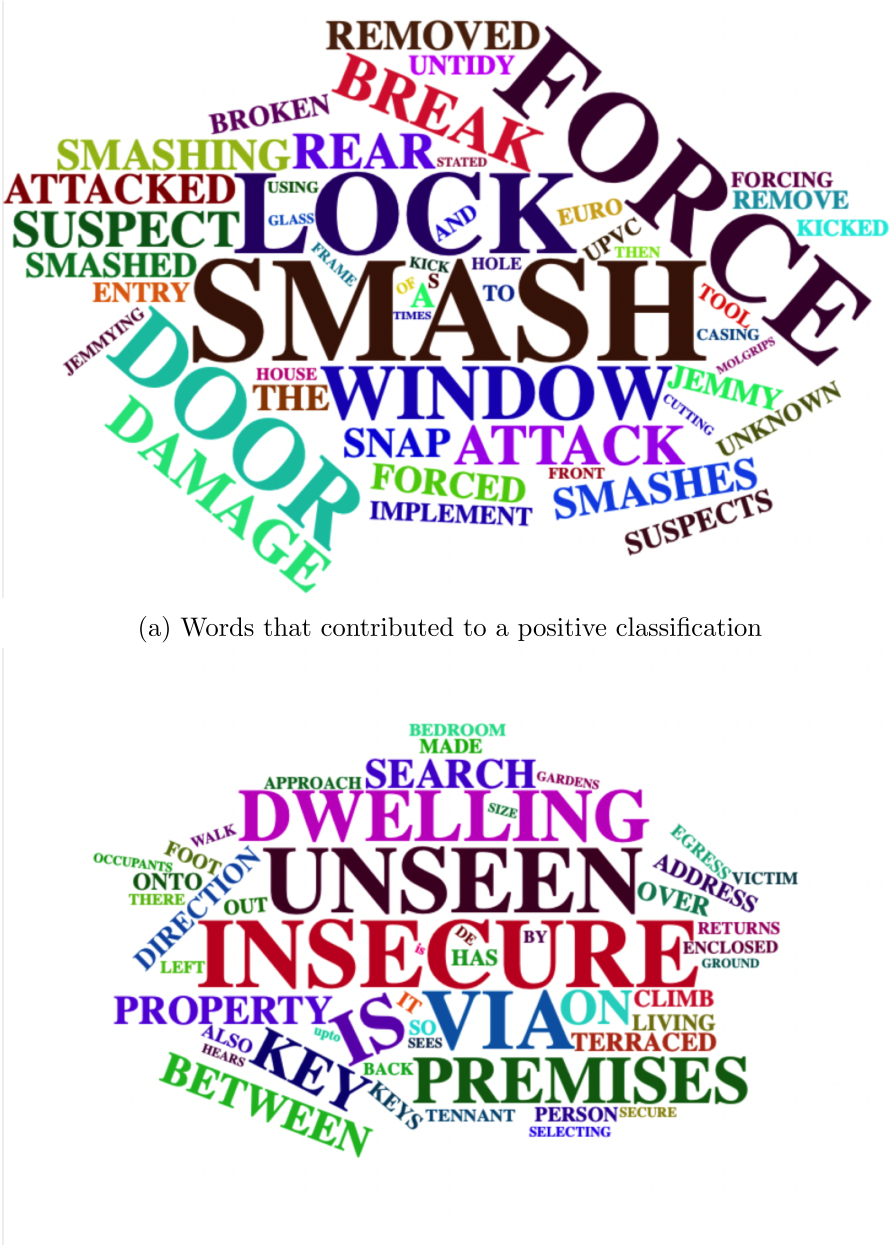
This section discusses the results that were presented in the preceding one. The section is structured around the questions that were outlined in Chapter 7.



(a) Words that contributed to a positive classification

(b) Words that contributed to a negative classification

Figure 10.2: Wordclouds from **motor vehicle** classification model. PF1 data. These wordclouds were generated using a fine-tuned BERT model on the PF1 data. The larger a word the more important it is for a classification. Words size is derived from a summation of the coefficients from individual LIME models. Word sizes are not comparable across figures.



(a) Words that contributed to a positive classification

(b) Words that contributed to a negative classification

Figure 10.3: Wordclouds from **force** classification model. PF1 data. These wordclouds were generated using a fine-tuned BERT model on the PF1 data. The larger a word the more important it is for a classification. Words size is derived from a summation of the coefficients from individual LIME models. Word sizes are not comparable across figures.

10.5.1 Can PTMs Classify MO Texts Accurately?

The results in Table 10.5 demonstrate that, in the limited classification tasks that are explored here, PTMs can classify MO texts accurately. High MCC scores indicate that the models learn the relevant patterns well and can classify unseen texts with a high degree of accuracy.

10.5.2 Are PTM better than the basic keyword method?

The results from the PTM and the keyword method are compared in Table 10.4. The keyword model and the PTM have similar recall values, and they both perform adequately when tasked with finding positive instances. The MCC values, however, are different. This difference shows that the PTMs are much more efficient on the whole. Although the keyword models find most of the positive instances of a classification, they also include many FPs. In other words, the keyword models classify more negative instances as positive than they should.

How problematic is this FP issue? The answer to that question depends on the problem and the number of texts that are overidentified. The absolute number of overclassifications might be manageable for rare classes because, even if one takes a large proportion of a small absolute value, the total number of false positives would be low. However, for more balanced classes, even the moderate overclassification of a large number of instances may result in the misclassification of a large absolute number of texts. It is this issue that affected the keyword model and the different classification problems, as explained in the section that follows. In the training data set, the number of MO texts that the keyword method labels as instances of theft of a motor vehicle is approximately 40% higher than the actual number of such thefts. Sometimes, for example, the MO describes the vehicle that was used to leave the scene of a crime, irrespective of whether it was stolen or not. Although the keyword search exhibits appropriate recall, in that it is likely to discover all of the burglaries that involved the theft of a motor vehicle, it also mislabels many other MO texts. Consequently, a thorough check of all labelled MO texts is necessary for a reliable labelling scheme to be produced. This requirement causes model building to take longer than previously argued. For example, there are 9,961 burglary texts, with an underlying base rate of vehicular theft of 9% (as estimated from the test and validation sets). The expected number of vehicular thefts within the burglary texts would therefore be 897. However, the keyword model classifies 1,727 texts as referring to a stolen vehicle. These 1,727 texts must then all be read for the FPs to be filtered. The estimated time that this task would consume is 17 hours, based on the earlier assumption of a reading speed of 100 texts per hour. Therefore, although the keyword model is much quicker to build initially (2 hours), approximating the performance of the PTM requires more time (19 hours).

The motor-vehicle classification that was presented above concerns an imbalanced data class. Accordingly, the keyword model could reduce the search space to a relatively small size, with an 80% reduction of the set of all burglary texts (from 9,961 to 1,727). However, since the force model concerns a more balanced classification problem (the estimated split between instances in which force is used and instances in which force is not used is 60-40), the keyword model cannot truncate the search space to the same extent as in the motor-vehicle classification problem. The use-of-force model can only shrink the search space by 17% (from 9,961 to 8,268) because the incidence of FP classifications is too high. For the remaining 8,268 texts, approximately 82 hours of labelling would be necessary to eliminate classification errors so as to approach the performance of the PTM.

The evidence indicates that the PTMs perform better than the basic keyword model when tasked with classifying MO texts. Although building PTMs and labelling texts takes longer initially, the results, as measured by MCC, are far superior than those of using the basic keyword models for a comparably shorter period of time. Two issues that have not been explored yet are that PTMs require specialist skills to operate and that keyword models can be made more intricate. That PTMs are complex is not in dispute. However, it is possible that they can be packaged for simple operation by nonspecialists so that there is no requirement to understand their complexities. However, at this stage, it would be unwise to dismiss the difficulty of implementing PTMs at police forces. This problem is discussed further in the final part of the thesis. Secondly, keyword models can be made more intricate, which would undoubtedly result in higher MCC scores. This said, PTMs emerged because probabilistic models were proven to be more robust than intricate rule-based ones – they are easier to maintain and generally exhibit superior performance when fed with unseen data.

10.5.3 Are PTMs explainable?

The evidence from the LIME models indicates that the PTMs use words that are consistent with human explanations for the classification of MO texts. This said, it is worth reiterating that the LIME models investigate local models of a selection of MO texts and that they are not explaining the model in a global context – not all words will have the same effect in every MO text.

The LIME output from the motor vehicle model (Figure 10.2a) shows that words such as “vehicle” and “car” are important for the classification of texts. This tendency is similar to that which humans exhibit and indicates that the model is operating similarly to a human tasked with classifying the texts. The model highlights meaningful words rather than ones with spurious correlations. The words that contribute negatively to the motor-vehicle classification are also of interest (Figure 10.2b). The words that are selected are more evenly sized and therefore of similar importance. Most of the words are common, which indicates that there are no particular patterns in the negation of the positive classification of motor-vehicle burglaries. This result was expected because past experiences with MO texts indicate that there are no instances of negative reporting. For example, no MO text states that a motor vehicle was not stolen. This, however, is not true of the use-of-force model. The application of force or its absence is generally noted in those texts.

Similarly to the motor-vehicle model, the LIME output from the use-of-force model is commensurate to the expected output of a human tasked with classifying one of the MO texts. “Force” and “smash” are prominent examples. In contrast to the motor-vehicle model, however, there is also a strong pattern in the words that contribute to negative classifications. The word “insecure” is prominent in the negative word cloud.

Although the explanations are local, the LIME output offers an adequate explanation of the classificatory choices. The “Explainable” section of the ALGO-CARE framework contains the following question: “Is appropriate information available about the decision-making rule(s) and the impact that each factor has on the final score or outcome?”. Arguably, the output is sufficiently explainable for the classification of each text to be justified. At the individual level, the texts are explainable, and a justification can be given for each classification by reference to the LIME explanations. At a global level, however, the model is not wholly explainable. If one takes the word “factor” from the quotation above to denote a word in the MO texts, then the factor in question cannot be said to have the same effect on the final classification in all instances. This difference results from the tendency of the model to use the surrounding context of a word as well as individual words to compute final effects. This tendency is problematic for all models in which strong interaction effects between factors are present, not just for text data.

The problem of explainability is unlikely to be settled on these pages. Tests would need to be conducted with a number of different stakeholders, including members of the public and police officers. However, the LIME output shows that the models work as expected, and classificatory decisions are made for appropriate reasons, that is, they are not based on spurious correlations.

10.5.4 Are PTMs Biased?

Models are biased if their performance differs systematically across types of instances. As noted previously, the PF1 is accompanied by limited metadata. There are no victim data on which to test bias. Therefore, only bias against text statistics is investigated here. Is there any bias that is related to the statistical properties of the texts, such as length or wording? The limited investigation did not detect any biases or systematic failings in relation to certain types of the text, as shown by the Pearson correlation coefficients. This finding has an important implication – if there is bias against certain victim characteristics which impacts, say, the length of an MO text, then that bias may not manifest itself as a degradation in probability accuracy. Therefore, a lack of correlation between victim characteristic and probability accuracy is not proof of absence of bias in the recording of the data, only of a lack of bias in model performance.

10.5.5 Limitations

The general limitations are discussed in the final part of the thesis. This section covers the limitations that are specific to the present study. In general, it has two main limitations. The first has to do with the number of classification tasks, and the second has to do with the investigation of bias:

- Classification Tasks. Only two classifications tasks were selected, and both are related to burglary. Although the two tasks concern problems of different types, that is, balanced and imbalanced classes, and although the results are satisfactory, the generalisability of the results to all types of crime is limited. The same is true of the applicability of the results to other classification tasks, such as the identification of burglaries that only target outbuildings. However, it is clear that there are problems for which PTMs can be a useful means of classifying MO texts.

- Bias. The bias investigation was severely limited by the lack of victim data within the data set. Study 1c compensates, for this shortcoming partially because it draws on victim data from which it is possible to assess classificatory bias.

10.6 Conclusion

In this narrowly focused study, PTMs were found to classify MO texts adequately across balanced and imbalanced classes. The tasks were to detect burglaries that involve the theft of a motor vehicle and burglaries in which force is used. In addition, the PTMs were found to perform better than simpler keyword models because they could discriminate more accurately in order to reduce the incidence of FPs. Despite the longer setup time and, in particular, the length of the process of labelling training data, PTMs are more efficient than keyword models. LIME was employed to understand how the models arrived at the classifications. In all cases, there appeared to be a sound rationale for the decisions of the models. The words that were more influential in the classificatory problems were also words that would have been important if a human had completed the classification tasks. Bias was only examined partially due to a lack of victim data. There did not appear to be any significant bias against texts of particular lengths or out-of-vocabulary words. This first study provides a solid basis for the use of PTMs. However, the use of active learning was not studied, and the applicability of the results is relatively narrow.

If applicability is to be expanded, it is important to discover whether the models work satisfactorily when used with other types of police free texts. Do they work when applied to different crimes? Only burglary was studied here. Do they work in other police areas, and can they facilitate the processing of other types of police data? These questions are answered, in part, in the subsequent chapters, in particular in the replication study that re-examines the classification problems that were explored here. That study concerns another police force. Before the replication study is presented, however, the next chapter investigates the use of active learning and its usefulness in reducing the burden of labelling.

Chapter 11

Study 1b: PF1 Active Learning

11.1 Introduction

The first study demonstrated that a large proportion of the user time that fine-tuning PTMs requires is taken up by the creation of labelled data from which the PTM can learn. Active learning was introduced in Chapter 4. It is a technique for reducing the labelling burden of training a model. Active learning is intended to reduce that labelling burden by highlighting the examples that are most helpful for improving the model. Active learning finds the texts that are most difficult to classify by using the PTM that has recently been fine-tuned to classify all unlabelled data. Once all of the data have been labelled, the instances with the closest individual classification probabilities are selected for labelling. This study explores active learning in order to determine whether its expected benefits materialise when it is employed with police data as well as the extent to which the additional processes slow the modelling process.

11.1.1 Problem Overview

This chapter concerns the fourth supporting objective - *Evaluate how effective active learning is with police data.* Effectiveness is judged by observing the MCC coefficient. If active learning is a better to an alternative (i.e. random) labelling strategy, its MCC score for an equivalent number of labelled data would be higher. This difference in MCC is also considered in view of the additional process that is required to enact the active learning strategy.

11.2 Active Learning Process.

Figure 11.1 depicts the general process of the active learning strategy. The first two steps of the process (the top left corner of the figure) are preparatory. Data are randomly selected and labelled for the test and validation sets. The third step is the final random selection. This third random selection yields 100 samples for the training set. Once selected, these samples are labelled by hand and used to fine-tune a model. The fine-tuned model is then used to predict the classification of all MO texts that are yet to be labelled. Once complete, the results of the model predictions are used to discover which of the MO texts the model was most uncertain about. Those texts are then labelled and added to the training set.

In practice, the output of the BERT model comprises log-probabilities for each potential classification, be it positive or negative. The absolute values of the differences in these log-probabilities are then ordered, and the MO texts that are associated with the 100 smallest values are selected. These 100 texts are labelled by hand to enable further fine-tuning. The train-predict-select cycle is repeated until it is decided that no further fine-tuning is necessary. At that point, the active learning process ceases.

The active learning process in this study was conducted in batches of 100. The number of 100 was selected because it results in an appropriate labelling time of around 1 hour. A longer process could have caused the concentration and the accuracy of the labeller to deteriorate. Selecting smaller batches may accelerated the convergence of the model predictions because the model would have adjudicated more often on the texts that were to be labelled. The benefits of convergence, however, would have been offset by the additional procedural overheads for each cycle, that is, the time that would have been necessary to find new data to label, to fine-tune the PTM, and to label all of the unlabelled data in accordance with the latest model.

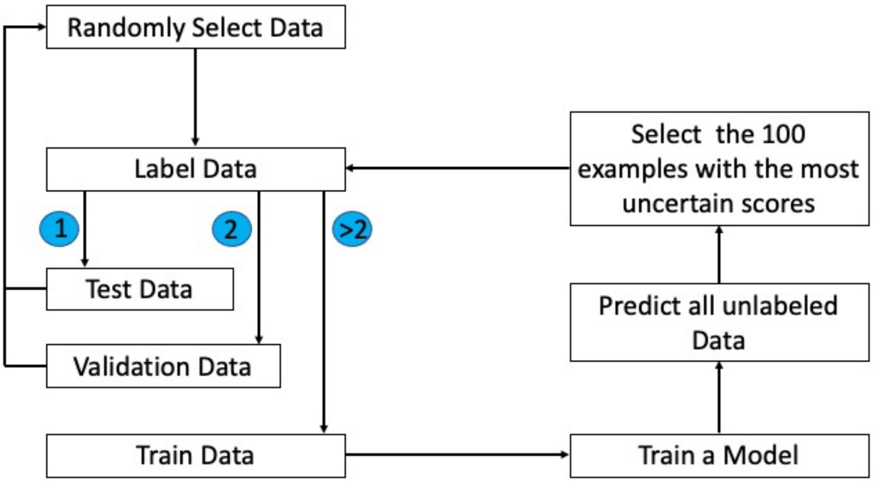


Figure 11.1: Active learning process, Step 1 and Step 2 – random labelling for test and validation sets. Subsequent steps entail using data that are initially labelled at random to train a model iteratively and selecting labels on the basis of model predictions.

11.3 Data and Method

11.3.1 Data

The data that are used in this chapter are the same as in Study 1a. They cover the burglary MOs from PF1. The classification problems are the use-of-force and motor-vehicle tasks from the preceding chapter.

11.3.2 Method

MCC scores are compared across models that are finetuned on data from the active learning process and models that are finetuned on pseudorandomly selected data. If higher MCC scores are obtained more rapidly with the active learning method, then active learning is assumed to have been beneficial. The difference in the number of batches that is required to reach an equivalent MCC score gives an indication of the utility of active learning.

The ideal method would be to compare the MCC scores from the active learning strategy with the MCC scores from the models that are based on a random approach to data selection. However, the random sampling approach was not adopted during the labelling of the MO texts. Instead, the active learning approach was compared to a pseudorandom sampling approach. Random sampling was not conducted in the course of data labelling because the available resources were insufficient to employ both the active learning technique and the random approach.

The pseudorandom sampling was generated by using the labelled text from an active learning approach that had not been applied to the model of interest. In this case, the data that were generated through active learning for the purposes of the use-of-force model served as a pseudorandom comparator for the motor-vehicle active learning approach. The following paragraphs focus on two issues that affect this approach and the manner in which they were investigated.

11.3.3 Potential Pseudo-random Problems

The first potential problem with the pseudorandom approach is that properties inherent in the MO text that make it difficult to classify. Accordingly, the pseudorandom approach may result in the selection of difficult-to-classify texts, regardless of the outcome, because it is based on an active learning strategy rather than on the correct active learning strategy. A random selection of data would not be of average difficulty because it would be truly random. Consequently, the perceived effect of the active learning strategy would be reduced. If this problem genuinely affects the data, there would be a significant overlap between the MOs that are selected by the use-of-force and the motor-vehicle active learning strategies.

In addition, and more importantly, there may be a correlation between the probability of selecting a positive use-of-force MO text and a positive motor-vehicle MO text through active learning processes. If this is the case, then the pseudorandom generation would be correlated with the model of interest. For example, the proportion of motor-vehicle labelled data in the force model would be higher than what one would expect from a truly random selection. This issue can be examined by comparing the proportion of active learning subjects (e.g. motor-vehicle theft) in the pseudorandom data (the data that are selected by using the use-of-force model). A proportion of active learning subjects that is close to expectations (i.e., the underlying random base rate) furnishes evidence against correlation. Such a finding can be verified by plotting the proportions of each classification in the selected data.

11.3.4 Pseudo-randomness Checks

The first check entails determining whether there is a large overlap between the MO texts that are actively selected for both models. To that end, a count of MOs that have been selected for the two models was completed. To be counted, an MO had to have been selected through the active learning strategy for both models in the first n selections, where n is the minimum of the two active learning pool sizes. A total of 22 MOs were selected for both models from a pool of 600. Therefore, 3.6% of the two active learning selections overlap. From this value, it may be inferred that the overlap is not large and that the inherent difficulty of the texts is not a significant factor in the results.

For the second test, which investigates the potential correlation between the two model types, plots that depict the proportion of each type of classification relative to the approach to selection are reviewed. These plots are displayed in Figures 11.2 and 11.3. The individual plots are explored below. Panel (b) of each plot is of particular interest for the second test.

Each plot has three lines. The red line denotes the expected random proportion, that is, the proportion that is calculated from the test and validation sets (400 samples). Assuming that they are randomly selected, this proportion should be an accurate estimate of the true-population proportion. For the motor-vehicle model, this proportion is 9%. The grey line denotes the cumulative proportion. It was calculated at each stage and for each sample, and its value is equal to the total number of positive samples divided by the total number of samples at a given stage. The black line is the batch proportion, that is, the proportion of positively labelled samples in a batch of 100.

Panel (b) plots the proportion of data with a positive classification for one model against the active learning selection of a different model. Therefore, Panel (b) in Figure 11.2 plots the proportion of positive classifications of theft of a motor vehicle with the active selection process of the use-of-force model. If the lines in Panel (b) are in close proximity, then the pseudorandom selection approximates a fully random one. The chapter now turns to a detailed exploration of the individual plots, which should indicate whether the pseudorandom data selection is a sufficiently close approximation of a random selection.

Motor vehicle plot We interrogate the lower plot of Figure 11.2 to see if the lines batch and cumulative proportions are close to the random proportion. Indeed we find that all lines are very close. This gives confidence that the data generated for the force model can be thought of as random with respect to the motor vehicle model.

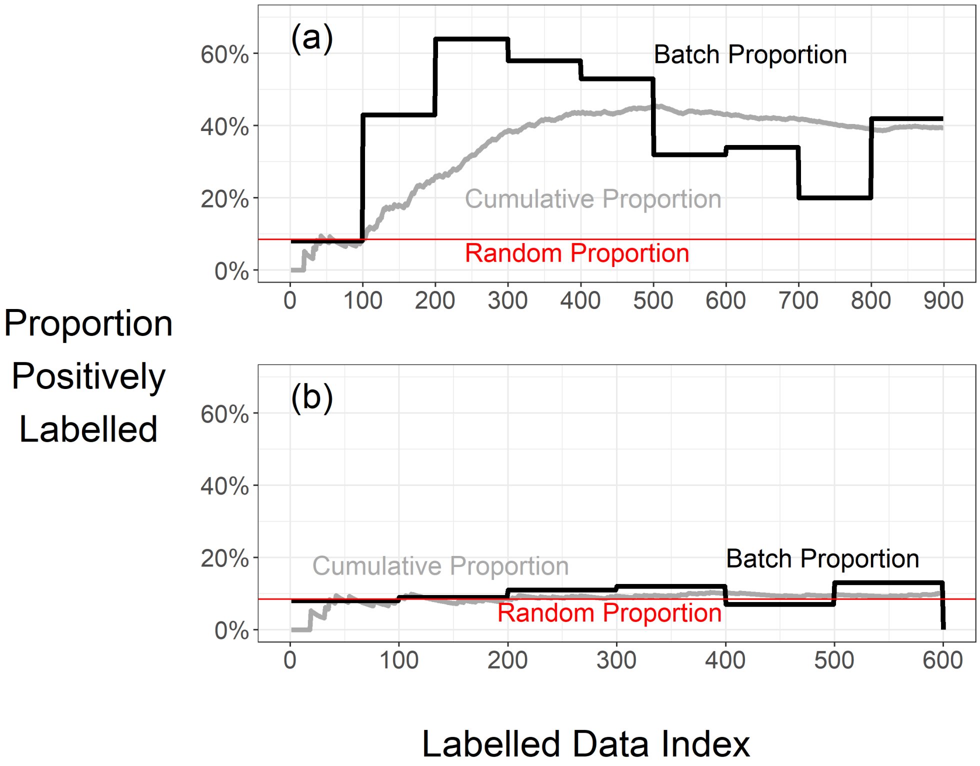


Figure 11.2: Active labelling - Motor vehicle model. The plots show the proportion of positively labelled burglary MO texts that were selected based on the theft of a motor vehicle classification. The labelled index (x-axis) indicates the order in which the data were selected and labelled. Panel (a) reflects the proportion of MOs selected that had a motor-vehicle stolen i.e in the second batch, 42% had motor vehicle stolen. Panel (b) This reflects the proportion of motor vehicle stolen MOs that were selected during the Force active learning.

Force plot The same procedure was employed to study the bottom plot in Figure 11.3. That plot covers the data that were selected by the motor-vehicle model but tested for the proportion of use-of-force cases. As with the motor-vehicle plot, the cumulative and the batch proportions are close to the random-proportion line. Once more, one can be reasonably confident that this use of the data represents a pseudorandom selection.

The results from the two plots and the investigation of the selection overlap indicates that the pseudorandom approach is sufficiently random to test the hypothesis that active learning would be an improvement on random selection to be tested. The next step entails comparing the MCC scores for each batch to determine whether the employment of the active learning approach has any benefits.

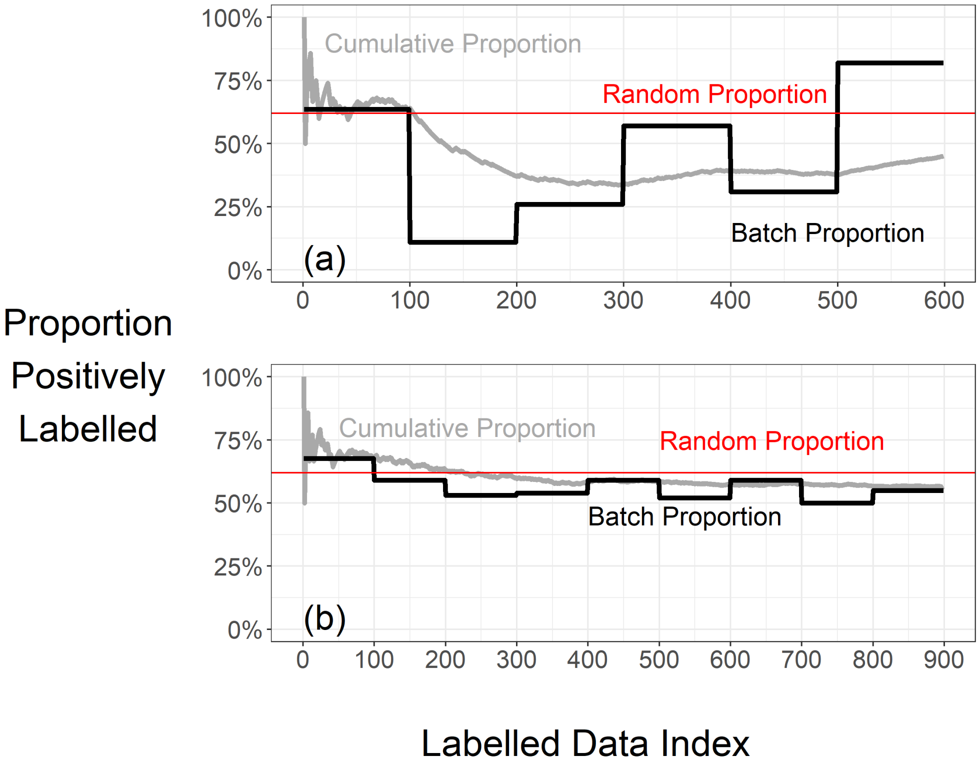


Figure 11.3: Active labelling - Force model. Plots showing the proportion of positively labelled MO notes for a burglary where Force was reported as being used. The labelled index (x-axis) indicates the order in which the data was selected and labelled. Panel (a) is for active learning based on Force model probabilities. Panel (b) is for selection using probabilities based upon the motor vehicle models.

11.4 Results

The results are explained for each model separately and then aggregated in the discussion section. The MCC scores for the active learning data and the pseudorandom data were compared as outlined above. The MCC scores were generated by using the validation dataset after each fine-tune. Recall that labelling ceased when the MCC of the validation set reached 0.9, with 1 being a perfect score. Active learning for the use-of-force model ceased after seven batches. Nine batches were needed for the motor-vehicle model.

11.4.1 Motor vehicle model.

Table 11.1 displays the MCC and recall metrics for each of the nine batches of active learning data. As expected, MCC generally increases as more data are labelled and peaks at 0.91 with 900 MO texts labelled. The final column in Table 11.1 displays the score that results from the use of all pseudorandomly selected data. This MCC score reflects the data that were generated from seven iterations of the use-of-force model. For a fair comparison, this MCC score should be contrasted to the seventh batch of the active learning-generated data. Active learning therefore has an MCC of 0.88, which is higher than the MCC value for the pseudorandom selection, which is 0.80. If one compares the score of the pseudorandom selection to all of the active learning values, it becomes evident that it falls between the scores for the fifth and the sixth sets. Therefore, the gain in model performance is equivalent to that of labelling 100 additional MO texts.

11.4.2 Force model.

The MCC values for the use-of-force model are presented in Table 11.2 . The last column of that table represents the MCC score for the seventh batch of the data that were generated from the motor-vehicle model. The final MCC of the active learning model is 0.92; the comparable MCC from the pseudorandomly generated data is 0.89. Once more, the active learning method produces a higher MCC score than the pseudorandom approach for a comparable number of labelled data. When compared to the MCC scores for active learning, the MCC of the pseudorandom selection falls between the fifth and the sixth set,

Batch	1	2	3	4	5	6	7	8	9	Force (7)
MCC	0	0.38	0	0.66	0.78	0.86	0.88	0.85	0.91	0.80

Table 11.1: MCC metrics for the motor vehicle model with the data selected through active learning. Each entry is the MCC metric after that batch. The final column refers to data that was selected using the alternative model (force model), the number seven in brackets refers to that data being the seventh batch and for most similar comparisons should be compared to the seventh active learning batch

Batch	1	2	3	4	5	6	7	Motor (7)
MCC	0.52	0.70	0.82	0.55	0.83	0.91	0.92	0.89

Table 11.2: MCC metrics for the force model with the data selected through active learning. The final column refers to data that was selected using the alternative model (motor vehicle), the number seven in brackets refers to that data being the seventh batch and for most similar comparisons should be compared to the seventh active learning batch

indicating a gain in performance that is equivalent to that which would result from labelling 100 additional MO texts.

11.5 Discussion

Active learning has been proven to be successful when used with police MO data. For both the use-of-force and the motor-vehicle model, the data that were selected through the active learning approach resulted in higher MCC scores than the pseudorandomly generated data. In essence, the benefit of active learning seems to be equal to that of labelling an additional 100 examples, a 14% decrease in the burden of labelling.

Clearly, the active learning approach has certain benefits. However, its use is not costless. Additional PTM finetuning is required. The model must be trained and allowed to label each of the texts. Exact training time varies with the time that it is allocated to finetuning and the generation of model predictions. For a deep learning model, this amount of time is not negligible.

For example, at the end of the active learning process, training the BERT model would take approximately 4 hours. A further hour would be needed to label the remaining data, contributing significantly to elapsed time.

Active learning also causes the process to become more complex. With several labellers, the process is delayed considerably because labelling must be co-ordinated in batches. This additional co-ordination period can also be lengthy. In practice, subsequent studies with multiple labellers showed that a single batch of texts is labelled in a 48-hour rhythm, which eases the burden on the labellers. What could have been achieved in a matter of days took a fortnight.

Additionally it is possible that the size of the batch (100) was too large. Perhaps labelling at a much reduced rate, say batches of 10 to 50, may have seen a greater reduction in overall labelling. This is because the model gets the opportunity to pick out the texts that it finds difficult to classify more often. Again though this reduction in batch size is likely to introduce proportionally more time lost to coordination and model building.

Therefore, the desirability of using active labelling is not self-evident. The decision must reflect a balance between the ease of adopting a more complex system and the time (both elapsed time and user time) that is available for a given task. If user time is limited, then active learning can save between 1 and 2 hours per project. It is also likely to result in more positive identifications of rare classes, providing the labeller with more exposure to MO texts of interest. However, if results must be obtained rapidly, that is, if elapsed time is of interest, and if the user can allocate more time to the task, active learning may be undesirable.

11.6 Conclusion

Active learning is a technique for reducing the amount of data that need to be labelled for the process of supervised learning to occur. The technique was used with PF1 data and applied to both the use-of-force and the motor-vehicle models. In both cases, the active learning strategy produced results that were superior to those that emerged in consequence of the adoption of a pseudorandom data selection strategy. However, the reduction in data labelling was only 14%. In practice, given the additional co-ordination costs that the active learning strategy entails, the resource savings are likely to be insignificant.

Chapter 12

Study 1c: PF2 Burglary MO

12.1 Introduction

The primary aim of this chapter is to examine the potential for replication of the work that has been presented so far. Replication is fundamentally important if work is to be undertaken on a large scale. Police forces in the UK can have different processes and training standards. Therefore, what works in one force area would not necessarily also work in another. The main rationale of this study is to provide further evidence for the purposes of Supporting Objective 2, "Evaluate how effective PTMs are with MO data".

In addition to replication, this chapter extends the analytic approach in four respects that reflect the different types of available data. First, the data allowed for an exploration of bias against victims with certain characteristics. Secondly, the data allowed for the completion of an additional classification task, namely an examination of burglaries that involve only entering an outbuilding. Third, it was possible to use the models that were developed for the PF1 data to label the PF2 data. This provides insights into the applicability of sharing the models across different police forces. Fourth, the models were fine-tuned on data from one year and used to label data from a subsequent year. This procedure sheds light on the decay of analytic performance over time.

The main finding of this chapter is that the results from PF1 are largely replicated with PF2, with no significant decline in performance. Accordingly, the key conclusion of this chapter is that PTMs are likely to be applicable at police forces other than the ones that are tested here.

12.1.1 Research Questions

This chapter aims to answer three research questions:

Can the results in Study 1a, the PF1 burglary study, be replicated in a different police force?

Study 1a inquired whether PTMs can be used to classify burglary MO texts in two different scenarios, which have to do with the use of force and the theft of motor vehicles. In each case, the PTMs were finetuned on PF1 data, resulting in appropriate accuracy. In this study, the PTMs are fine-tuned on the same problems but with PF2 data. In addition, it was also possible to build a model for the outbuilding only model. The outbuilding only model was introduced in Study 1a, but it was not completed because the PF1 data did not contain references to that type of burglary. The outbuilding only model is intended to detect whether a burglary targeted only an outbuilding, such as a shed, without the main home of the victim being breached.

Can models trained with data from one police force be used in another force?

Fine-tuning PTMs on the same task at two different police forces enables the models to be used across areas. It also becomes possible to ascertain whether they are generalisable. If their applicability is indeed broad, then large benefits are likely to result from the dissemination of the models across police forces, which would reduce the resource burden of model creation. This problem is outside of the scope of the first question, which presupposes that the models are built from data for a particular police force. In this second question, the models are built with PF1 data then used on PF2 data.

Are models accurate over time?

Language changes both through the introduction of new words and as a result of changes in the usage of existing lexical units. In policing contexts, officers can also be encouraged to record different facts over time. These changes could potentially change the form or the wording of an MO text. If the language of MO texts changes so as to differ from the language that the PTMs are finetuned on, then one can expect performance to deteriorate. Although only two years' worth of data are used here, this hypothesis is tested by finetuning a model on data from one period and testing it on data from a subsequent period. Understanding how or when the performance of a model may deteriorate is important for ensuring that the model that is being used has been trained correctly.

12.2 Data

The data that are used in this study are from PF2. They were described comprehensively in Chapter 8. The PF2 data were whitelisted by the project team in order to remove personally identifying information. This process was also described in Chapter 8. The main difference from the PF1 data results from the addition of victim characteristics, namely ethnicity and sex, as metadata. This addition is conducive to a more profound investigation of the potential biases that the PTMs produce.

Beyond victim characteristics, additional details were provided after the models had been built for validation purposes. Links to stolen vehicles were added in order to facilitate the validation of the vehicular theft model. A link is an entry in the police database that indicates whether a vehicle was stolen during a given burglary. This provides an additional verification that enables the performance of the finetuned PTM to be assessed. The completion of the database link results in structured data that is easy to search. PF2 analysts expect “stolen” links to have higher completion rates than flags (the structured data that were introduced earlier). A “stolen” link is therefore an appropriate structured indication of whether a vehicle was stolen. It can be compared to the model classification of the text data.

The PF2 data also contained more details about the date and time of the offences that had been committed. The PF2 data included details on the years, months, days, and times of offences, whereas the PF1 data included only months. Consequently, the data on dates can be analysed in greater detail. As an interesting aside, the PF2 data also cover the period of the initial Covid-19 pandemic in the UK, including the first lockdown. The effects of that lockdown on intra-crime variation for burglary can also be observed.

12.3 Methods

The methods of this study were introduced in Chapter 9 and recapped in Chapter 10, which is on Study 1a. The general process is similar to that which was explained previously. The deviations from the approach that was described in the general introduction are listed below.

12.3.1 Labelling

As in Study 1a, the fine-tuning of the PTMs is a supervised learning process. Therefore, labelled data are required for the models. The data were labelled by two researchers, with the author holding the casting vote in the event of disagreement. The MO text was selected through an active learning strategy, as detailed in Chapter 9. On this occasion, the labelling data pool was limited to burglaries committed between October 2018 and the end of 2019 (as mentioned in the data chapter, October 2018 coincides with the introduction of the new data-recording system for PF2). This restriction was introduced in order to facilitate the investigation of the accuracy of the model over time, that is, to enable the third research question of the study to be answered. Active learning was only conducted for the motor-vehicle model (reason explained later). Accordingly, all PTMs were fine-tuned on data that had been selected for the motor-vehicle model through active learning. In total, 1,982 MO texts were read and labelled for the burglary classification models.

Fine-tuning Models

There were no significant differences between the fine-tuning methods that were applied to the PTMs in this study. Fine-tuning was completed by using the same methods as those that were outlined in Chapter 9. The BERT-large model was used. The hyperparameters were all set in the manner that is described in Chapter 9.

Performance

The additional data fields that are provided with the PF2 data allowed for a deeper investigation into bias than had been possible with the PF1 data. Bias against individuals of certain sexes and ethnicities was explored by comparing PTMs across different victim characteristics. Bias was explored by using the metrics Equality of Opportunity (EoO) and Predictive Parity (PP), both of which were introduced in the methods chapter. EoO is based on recall and measures the disparity of the probability of a true positive (TP) across groups. For example, given that a classification is positive, what is the probability of finding it? PP is based on precision and is a measure of the disparity of the probability of false positives (FPs) across groups. For example, given that the model finds that a classification is positive, what is the likelihood that the classification is correct?

These metrics were calculated for each test set, and a cross-validation experiment was completed. As noted earlier, a reference group was selected for each bias in order to determine whether there is a difference between the reference group and the remainder of the population. The reference groups were "white European" and "male", and they were compared to the groups "all other ethnicities" and "females", respectively. Unknown and missing values were excluded from the analysis. [1].

No comparison to a basic keyword model was conducted. The advantages of the PTMs over the keyword approach were explained in Chapter 5 and demonstrated in Study 1a. However, it was possible to make a comparison with another method that police forces may use. Police forces often record some aspects of intra-crime variation as flags. Flags are typically key words or phrases that a police officer can select to describe a crime. In the PF2 data, these flags had been selected from a series of dropdown menus on the crime-recording software. These flags are much easier to search than free-text data because they are structured and are therefore be used often to find

[1] The analysis was also conducted with these missing values included in the comparison groups and there was no significant difference in the result.

Classification	Flags
Motor vehicle	Instrument Used, Key Used, Stolen Instrument Used, Key Used, Key Used Property, Conveyance, Car Property, Conveyance, Motorcycle
Force	Trademarks- Attack Method Premises Entry Method, Attack Method Premises
Outbuilding	Location, Garage - Includes premises for sale and repair but does not include petrol station Domestic—location, Garden - Driveway, Shed

Table 12.1: The keywords used to filter the MO Keywords data column in the PF2 Burglary data.

crimes of interest. The following process was followed in order to compare the flags to the model: firstly, a decision was made about the flags that describe classification types accurately. For example, it was determined which flags highlight burglaries in which force was used. The list of flags that were used for each classification is displayed in Table 12.1. Secondly, the monthly counts of crimes that do and do not meet the criteria of the classification were summed. This step was completed for both the crimes that were selected by reference to flags and to the crimes that were selected through the use of the NLP model. It was possible to compute monthly percentages of positive classifications from the sums of the positive and the negative classifications. Finally, the percentage of positive classifications was plotted as a time-series line, and the line plots were compared.

As mentioned in the data section, there was an additional validity check on the motor-vehicle classification, namely for the presence of a link between a stolen vehicle and the crime. These data were also used to check the validity of the vehicular theft model and were added to the monthly time-series plots that were described in the previous paragraph. The calculation method was the same.

Model Performance Over Time.

Model performance over time was investigated by fine-tuning the model on data from one period and testing it on data from a subsequent period. All PTMs were fine-tuned and initially tested on data from the period between late 2018 and the end of 2019. For the replication element of this study, all MCC metrics were gathered from a test set that was randomly selected from the same set of dates. However, a separate test set was also built for 2020, which allowed the model from the earlier time period to be tested on the 2020 data. It is possible that the 2020 data are not ideal for comparative purposes due to the Covid-19 pandemic. The pandemic resulted in severe mobility restrictions as the government tried to curtail the spread of the virus, and it expanded the lexicon of the general public. However the effect of the pandemic on burglary MO texts may have been less severe because it is not immediately clear how the pandemic would change burglary methods and, therefore, the words that the police use to describe burglaries. However, if there is a significant degradation in model performance over the years, then the Covid-19-induced variation would be one source of change that would require further investigation.

PTM transfer-ability

PTMs that were fine-tuned on PF1 data were used to label the PF2 test sets. MCC scores were calculated and directly compared to the MCC scores from the models that were built from the PF2 data. For example, the force model that was built from the PF1 data was used to label the PF2 test set, and the labels that were generated were compared to the force labels. Comparing the two MCC scores indicates how accurate a model that is generated by one police force can be when it is employed by another police force.

12.4 Results

The results of the replication study are presented first. The MCC scores and explainability are directly comparable to the earlier study of the PF1 data because the two studies are based on the exact same methods. The estimates of bias are different because the PF2 data cover victim characteristics. Therefore, the metrics of extrinsic bias for the sex and ethnicity groupings were examined. The comparison with the NLP labels for which the police recorded keywords are displayed as line plots after the bias results. Thereafter, the exposition turns to the MCC results for the change in time period and the reuse of models across forces.

12.4.1 MCC

The MCC results are presented in Table 12.3. The table refers to the test sets from 2018–2019 and from 2020–2021. However, before these sets are explored, it is necessary to explain, in brief, how much data were labelled and why.

As detailed in the methods chapter, when the active learning strategy was used, labelling for the validation set would cease when MCC exceeded 0.9. The motor-vehicle model, which was selected for labelling first, never achieved this value (see Table 12.2). For the reasons that are given in the next paragraph, additional labelling was unlikely to result in increases in MCC. Consequently, all labelling ceased after the 16th active learning batch.

The motor-vehicle task was selected for labelling first. However, all tasks (i.e. the use-of-force and outbuilding only tasks) were labelled at the same time. At Batch 16, enough texts had been labelled to gauge the necessity of additional labelling for the two other classification tasks. The use-of-force classification task had achieved an MCC of 0.92 by Batch 5 (see Table 12.2). Therefore, no further labelling was necessary.

The outbuilding task did not reach an MCC of 0.9 by Batch 16; the MCC scores stabilised at approximately 0.85 at Batch 5 (see Figure 12.1.). It was therefore unlikely that additional labelling would increase the MCC score. However, a single batch of additional active learning was conducted, with a fine-tuned outbuilding model applied to the final selection. The MCC of the model (tuned on 16 batches of motor-vehicle and one batch of selected outbuilding data) did not increase as a result. Therefore, the author determined that no further labelling was necessary. The data from this 17th batch are omitted for simplicity.

The MCC scores for the models are reported in the subsections that follow. An MCC score of 1 is optimal, while a score of 0 is equivalent to a finding of

Batch	1	2	3	4	5	6	7	8
Motor vehicle	0	0.48	0.65	0.73	0.56	0.82	0.8	0.82
Force	0.52	0.71	0.82	0.88	0.92	0.93	0.88	0.92
Outbuilding	0.33	0.23	0.72	0.84	0.87	0.85	0.85	0.86
Batch	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Motor vehicle	0.75	0.72	0.88	0.86	0.82	0.75	0.82	0.72
Force	0.88	0.92	0.92	0.9	0.92	0.91	0.93	0.91
Outbuilding	0.85	0.86	0.85	0.84	0.86	0.84	0.85	0.86

Table 12.2: MCC values (based on the validation set) for models fine-tuned on PF2 Burglary data. Batch refers to the active learning batch e.g. after 5 batches of labelling (500 MO texts) the motor vehicle model had an MCC of 0.56

randomness. The results that are discussed below are confined to the 2018-2019 test set. The 2020-2021 test set scores are discussed at a later stage.

Motor Vehicle model

As explained previously, the motor-vehicle model was selected as the first model for labelling via the active learning strategy. The values of MCC after each active learning batch, which were calculated by using the validation data, are displayed in Figure 12.1 and in Table 12.2. The highest value that was attained was 0.88, which is below the requirement of the stop condition (0.9). Labelling ceased after the 16th batch because it had become apparent that there were no further positive classifications within the pool of potential training data. In other words, the active learning strategy had already selected all MO texts that refer to the theft of a motor vehicle into the training data, and there were no more positive examples that could be used for learning. In fact, the last positive example had been found in Batch 11. It is clear from the plot in Figure 12.1 that the additional negative examples, which formed Batches 12–16, did not facilitate the fine-tuning of the model. Therefore, further fine-tuning was

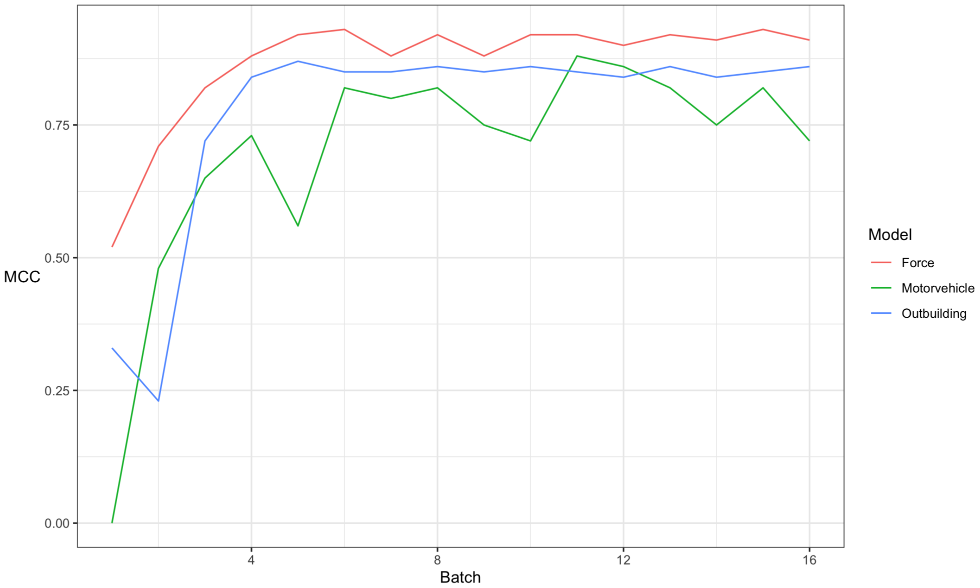


Figure 12.1: MCC scores for the PF2 burglary models. MCC scores are shown after each iteration of the active learning strategy. The Force and Outbuilding models peak relatively early on at batch 5 and 6. Whereas the motor vehicle model peaks at 11. Some of the variation will be attributable to the random initialisation of the models. Source: Author generated.

deemed unnecessary. Consequently, the fine-tuning stopped after 16 batches, and the model was tested on the test set.

The MCC scores for the test set can be found in Table 12.3. The model that was finetuned over 10 runs had a mean MCC score of 0.98, which is indicative of near-perfect performance. In half of the runs, the model classified each of the 200 MO texts in the test set correctly. The MCC metrics for the motor-vehicle model are comparable to the scores from the motor-vehicle model that was built on and applied to the PF1 data (mean of 0.97).

Force model

The use-of-force model was applied to the same data that were labelled during the active learning for the motor-vehicle model. The mean MCC score from the 10 final initialisations on the 2018-2019 test set was 0.93. These scores are higher than the MCC score for the validation set, indicating that the validation set may have made the classification of the MO texts more difficult. These MCC results are comparable to the models that were finetuned and tested on the PF1 data (mean of 0.91).

Outbuilding Model

The MCC scores for the application of the outbuilding model to the test data are also better than the validation set scores. The mean of the 10 initialisations on the entire 2018-2019 test set is 0.90. This is the lowest score across all three models. The outbuilding model was not built with the PF1 data because they are not suitable for its purposes. Therefore, the outbuilding results that are presented in this study cannot be replicated directly.

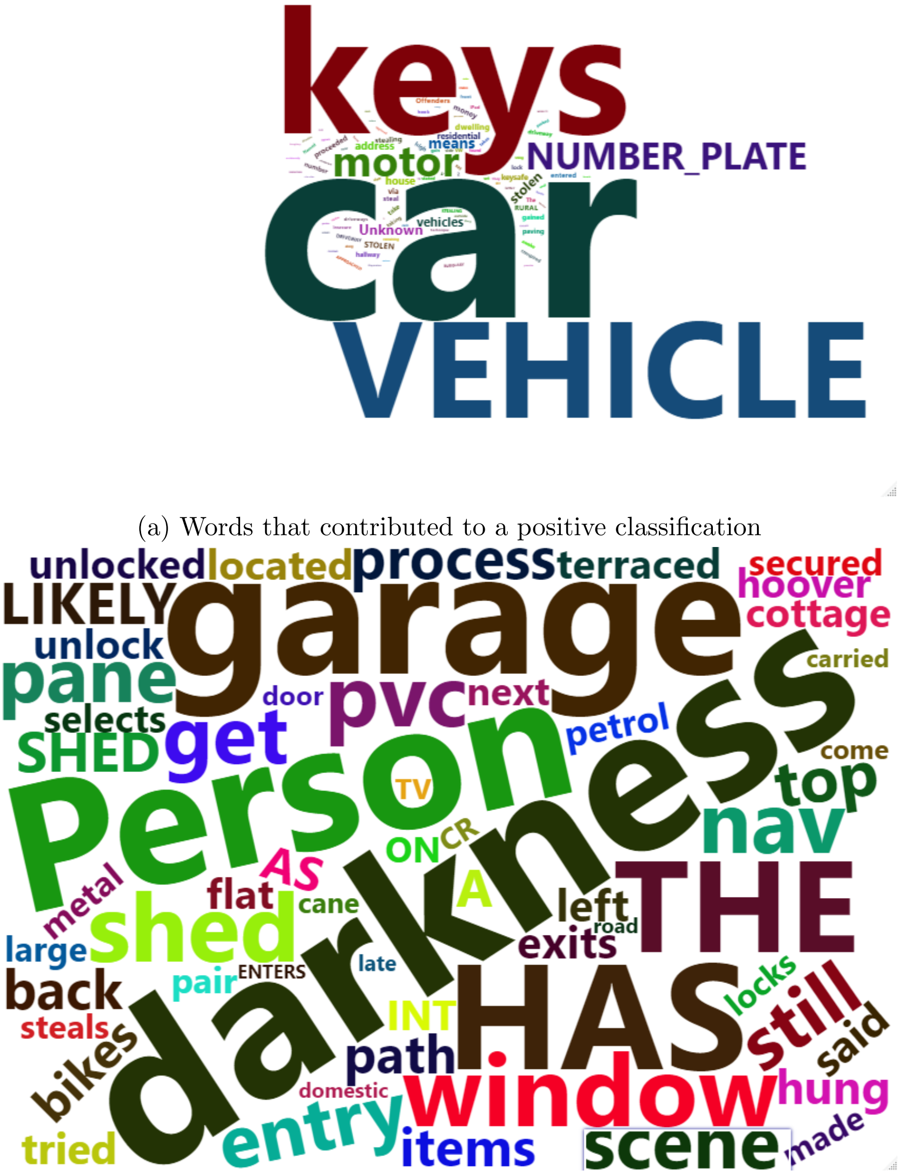
12.4.2 Explainability

LIME was used once more to understand how the words in the texts contribute to the final classification. By way of reminder, BERT uses words and their surrounding context. Therefore, it is difficult to form a global understanding of the workings of the model. LIME provides a local understanding of each MO text by deleting words randomly to enable their effect on the final classification to be discerned. This approach is scaled up in this thesis through the application of LIME to all MOs in the test set and through the use of word clouds to highlight the most important words, as identified by the individual LIME model coefficients. The word clouds for the motor-vehicle, use-of-force, and outbuilding models are displayed in Figure 12.2 , Figure12.3 and Figure 12.4 respectively.

The word clouds for the motor vehicle model exhibit a similar pattern to those from the PF1 data, see Figure 12.2. Firstly the most prominent words in the word cloud for a positive classifications (i.e., a motor vehicle was stolen) are words that a human might expect to use when completing the same classification task. The three most important words are “car”, “vehicle”, and “keys”. It should also be noted that these words are disproportionately important, which is why their size in the figure is much larger than that of other words. In contrast, the word cloud for words that contribute to a negative classification (Panel B in Figure 12.2) contains words that are much more similar in size. There is no observable theme, likely because the absence of car theft is not explicitly recorded in the MO texts.

	Motorvehicle		Force		Outbuilding	
Run	18/19	20/21	18/19	20/21	18/19	20/21
1	1.00	0.89	0.93	0.92	0.91	0.94
2	1.00	0.93	0.94	0.93	0.90	0.93
3	0.94	0.88	0.93	0.94	0.90	0.94
4	0.97	0.93	0.93	0.95	0.89	0.93
5	0.94	0.88	0.94	0.96	0.92	0.94
6	1.00	0.91	0.90	0.93	0.91	0.94
7	0.97	0.88	0.90	0.94	0.85	0.96
8	1.00	0.91	0.92	0.95	0.90	0.92
9	0.97	0.90	0.92	0.96	0.91	0.94
10	1.00	0.90	0.95	0.96	0.89	0.92
Mean	0.98	0.90	0.93	0.94	0.90	0.94
Best Run	1.00	0.93	0.94	0.96	0.92	0.96

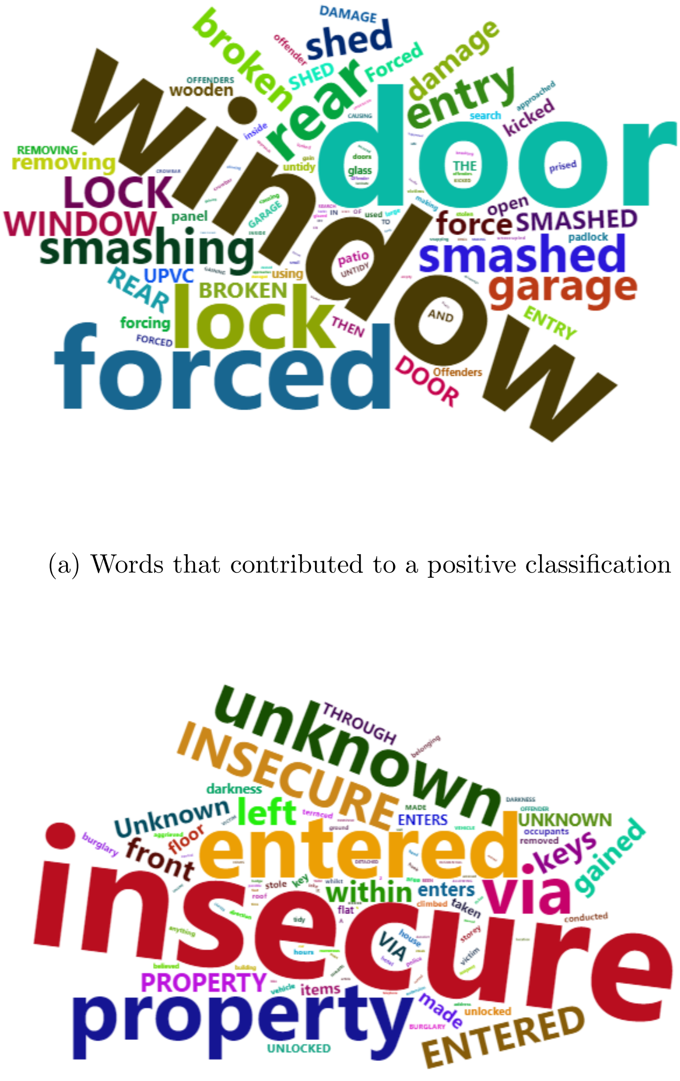
Table 12.3: MCC values (based on the test sets) for models fine-tuned on PF2 Burglary data. Scores are generated from 10 separate fine-tunes based on all labelled data. 18/19 refers to the test set from only the years 2018 and 2019, similarly 20/21 refers to the years 2020 and 2021.



(a) Words that contributed to a positive classification

(b) Words that contributed to a negative classification

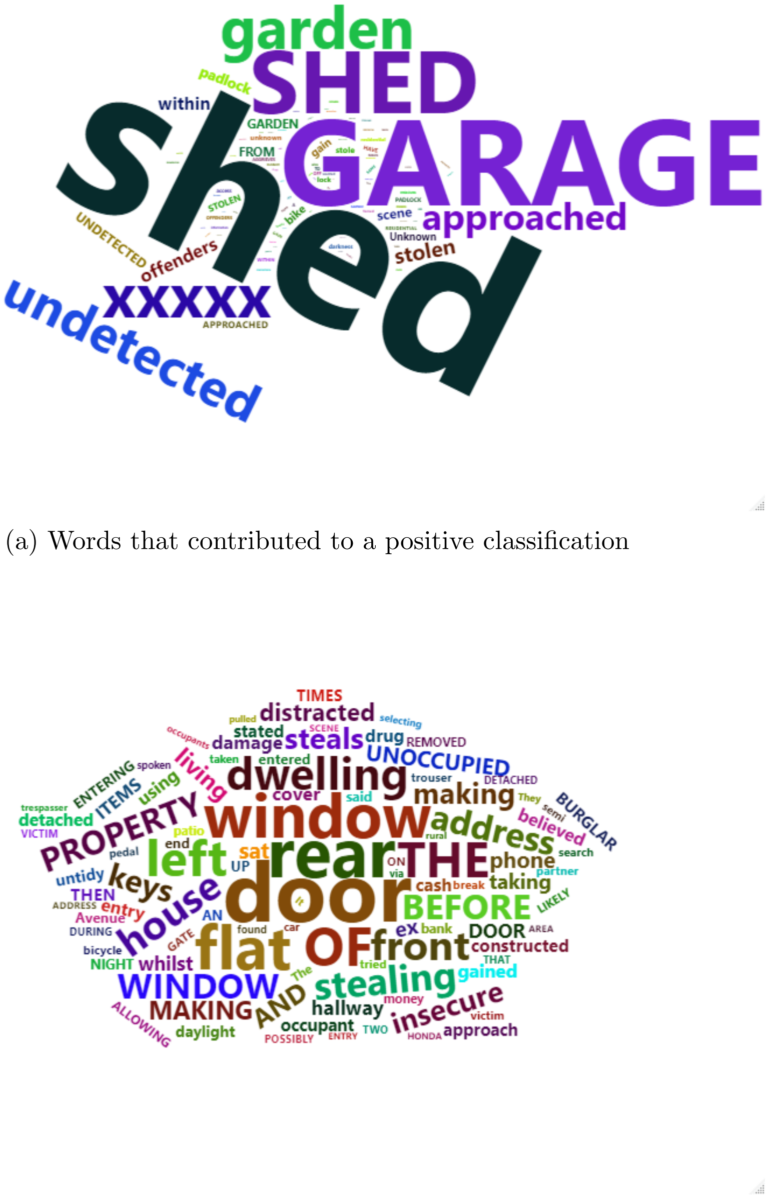
Figure 12.2: Wordclouds from **motor-vehicle** classification model. PF2 data. These wordclouds were generated using a fine-tuned BERT model on the PF2 data. The larger a word the more important it is for a classification. Words size is derived from a summation of the coefficients from individual LIME models. Word sizes are not comparable across figures. Source: Author generated.



(a) Words that contributed to a positive classification

(b) Words that contributed to a negative classification

Figure 12.3: Wordclouds from **force** classification model. PF2 data. These wordclouds were generated using a fine-tuned BERT model on the PF2 data. The larger a word the more important it is for a classification. Words size is derived from a summation of the coefficients from individual LIME models. Word sizes are not comparable across figures. Source: Author generated.



(a) Words that contributed to a positive classification

(b) Words that contributed to a negative classification

Figure 12.4: Wordclouds from **outbuilding** classification model. PF2 data. These wordclouds were generated using a fine-tuned BERT model on the PF2 data. The larger a word the more important it is for a classification. Words size is derived from a summation of the coefficients from individual LIME models. Word sizes are not comparable across figures. Source: Author generated.

The word clouds for the use-of-force model are similar in the positive classification case (i.e., force was used). The model uses words that are similar to the ones that a human might use if entrusted with the same classification task. However, some of the more important verbs are less prominent, and there appears to be a stronger focus on nouns, in comparison to the word clouds for the PF1 data. On the whole, the pattern of the important words is clear and logical. Unlike the motor-vehicle model, the negative classification, “no force used”, is often reported, and there is a clear pattern in the second word cloud (Panel B in Figure 12.3). The word “insecure” is the most important, for obvious reasons. “Unknown” is also prominent because it is often used to indicate that the method of entry is unknown, reflecting lack of clear evidence of use of force.

The outbuilding word cloud is similar in structure to the motor-vehicle word cloud. The positive classification cloud contains a smaller number of disproportionately important words. “Shed”, “Garage”, and “garden” are the most important among them. The negative classification word cloud contains words that are closer in size and, in general, words that are encountered throughout all burglary MO texts. Again, this tendency is likely the result of failure to report on negative classifications explicitly.

Each of the pairs of word clouds indicates that the model uses words that a human would also rely on in determining the classification of a text. It may thus be inferred that the models focus on the most appropriate features of the text and not on spurious correlations. The next section investigates the biases that the models may exhibit in relation to sex and ethnicity.

12.4.3 Bias

Bias within the models was investigated by reference to the characteristics “sex” and “ethnicity”. The reference groups were “males” and “white Europeans”. The models were investigated by exploring metrics of extrinsic bias, EoO, and PP. Table 12.4 displays the results from the two models that were built from the active learning data and from the tenfold cross-validation experiment models. The results are described in relation to the partition of the data, that is, by reference to sex and ethnicity rather than to specific models. The reader will recall that 0 is indicative of no bias, that a positive number is indicative of bias in favour of the reference group, and that a negative number is indicative of bias against the reference group. The theoretical maximum and minimum values are 1 and -1, respectively.

Ethnicity

Two significant p values emerged from the cross-validation experimentation, and they are both for ethnicity. One is for EoO in the motor-vehicle model, and the other is for PP in the use-of-force model. In both cases, the mean shows that there is a slight bias against the reference group, that is, that the models may discriminate against white Europeans.

A review of the results from the model that was built from the active learning data indicates that most values for EoO and PP are close to zero, indicating little bias. Four out of the six EoO metrics are negative; the same is true of five of the six PP metrics. Once more, both findings indicate that there is slight discrimination against white Europeans. The results across the models are mixed. The direction of the bias is only consistent in the outbuilding model. However, even then, the bias in that model does not produce a statistically significant result in the cross-validation experiment. In summary, the values that reflect bias are small, but the results are not sufficiently consistent to indicate that the PTMs that classify MO texts in the PF2 data are systemically biased.

Sex

The evidence for bias on the basis of sex is weaker still, and no statistically significant results emerged from the cross-validation experiments. For PP, the number of negative values is equal to the number of positive values. For EoO, the number of positive values exceeds the number of negative ones by one. In conclusion, there is no evidence that the PTMs that classify MO texts from the PF2 data exhibit bias against either sex.

Equality Of Outcome

Model	Partition	Test set 18/19	Test set 20/21	CV Mean	CV p value
Motor vehicle	Ethnicity	0.000	-0.040	-0.049	0.001*
Motor vehicle	Gender	0.000	-0.129	0.002	0.902
Force	Ethnicity	-0.022	0.056	0.004	0.593
Force	Gender	0.048	0.014	0.004	0.530
Outbuilding	Ethnicity	-0.012	-0.017	-0.005	0.462
Outbuilding	Gender	-0.017	0.001	-0.004	0.147

Predictive Parity

Model	Partition	Actual 18/19	Actual 20/21	CV Mean	CV p value
Motor vehicle	Ethnicity	0.000	-0.077	0.155	0.078
Motor vehicle	Gender	0.000	-0.005	0.045	0.060
Force	Ethnicity	-0.040	-0.040	-0.108	0.024*
Force	Gender	0.091	-0.120	0.001	0.948
Outbuilding	Ethnicity	-0.012	-0.017	-0.003	0.821
Outbuilding	Gender	-0.032	0.009	0.000	0.989

Table 12.4: Extrinsic bias metrics for the Lancashire Burglary models. AL refers to the model built with data selected by active learning, the following digits represent the year of the test set. The mean refers to the mean result from the 10 cross-fold validation experiment. The p value relates to the hypothesis test that the mean, from the cross-fold experiment, is not zero. * is for a p value that is significant.

12.4.4 Flag Comparison

This section compares the NLP-generated labels for the three models with the flags that the police may search in order to identify an intracrime variation of interest. In addition, the presence of links between stolen vehicles and burglary is explored.

Motor Vehicle Model

The time-series plot for the motor-vehicle model is displayed in Figure 12.5. The police-generated labels “Linked vehicle” and “Flagged” should only be fully considered after 2019 because of the aforementioned change in data-recording systems. The two striking elements of the plot are that the NLP labels and the linked-vehicle labels are very well matched (the Pearson correlation coefficient is 0.94) and that the flags cover much fewer crimes. The latter tendency is observed across the classifications. In a discussion, the analysts from PF2 recognised that the flags do not have a high completion rate. However, based on their experience, they thought that the linked-vehicles data would be completed to a high standard.

An error analysis was conducted in order to explore the differences between the NLP model and the linked stolen vehicles. The error analysis reviewed 100 MO texts in which the NLP model had identified references to a stolen motor vehicle but for which there was no linked vehicle. In total, there were 432 errors of this kind. Of the 100 MO texts that were examined, 63 did refer to a stolen vehicle. Therefore, the classification from the NLP model was correct. The remainder (37) had been labelled incorrectly by the PTM. The majority of these errors (21) occurred when only vehicle keys had been stolen. These erroneous classifications may be useful in the context of vehicular theft because keys can be used to steal a vehicle after a burglary, but they do not reflect the purpose for which the model was trained.

Force Model

The use-of-force model only compares flags to PTM labels. The plot is displayed in Figure 12.6. As with the previous time-series plots, the most notable finding

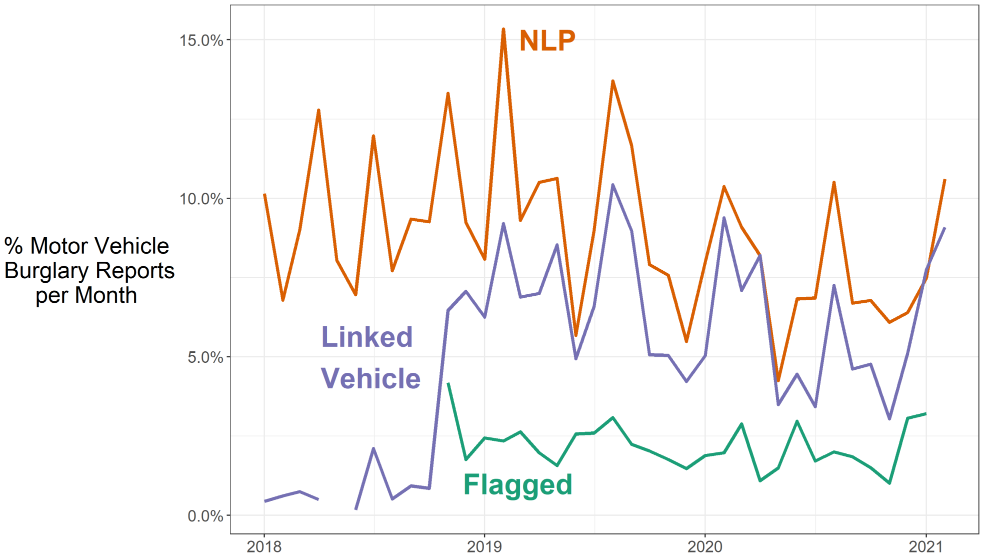


Figure 12.5: A time series plot of the motor vehicle classification. Showing data generated form the PTM model (NLP), linked vehicles and flags. Source: Author generated.

is that the PTM finds many more burglaries that involved the use of force than the police officers who use the flag system. Once more, this finding is consistent with the analysts' view that the flag system is not used appropriately in practice. The other notable finding is that the PTM labels appear to be seasonable – the proportion of crimes in which force is used to enter a building is consistently higher in the winter months than in the summer months.

Outbuilding Model

The outbuilding model plot displays the PTM-generated labels alongside the police-generated flags. The plot is displayed in Figure 12.7. Like in the case of the other two plots, the PTM returns more crimes with a positive classification. However, the numbers that are returned here are closer than in the other plots. The early 2020 spike in the two time series coincides with the first Covid-19 lockdown in the UK. This spike may indicate that the lockdown policies resulted in a proportional shift in burglary types.

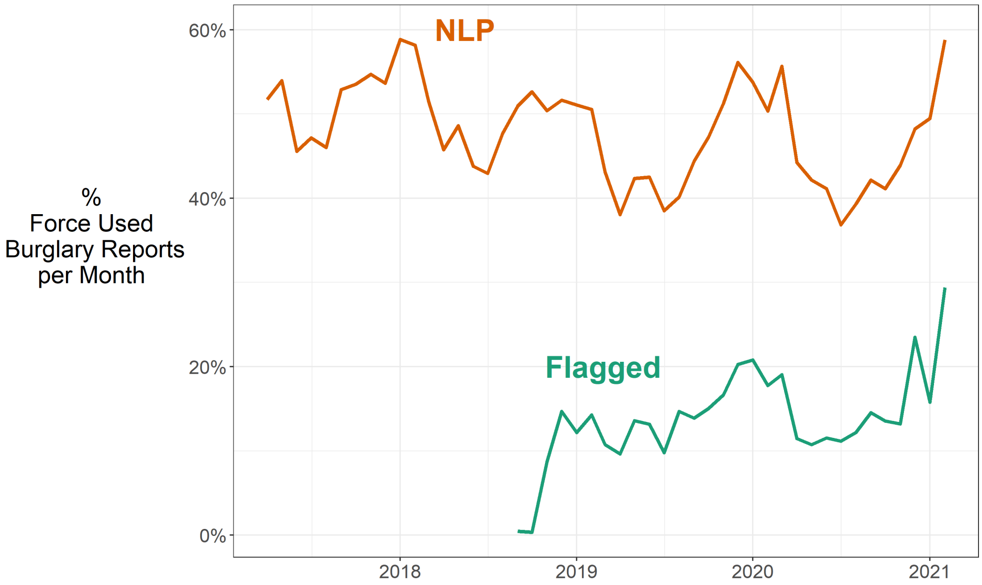


Figure 12.6: A time series plot of the force used classification. Showing data generated form the PTM model (NLP) and flags. Source: Author generated.

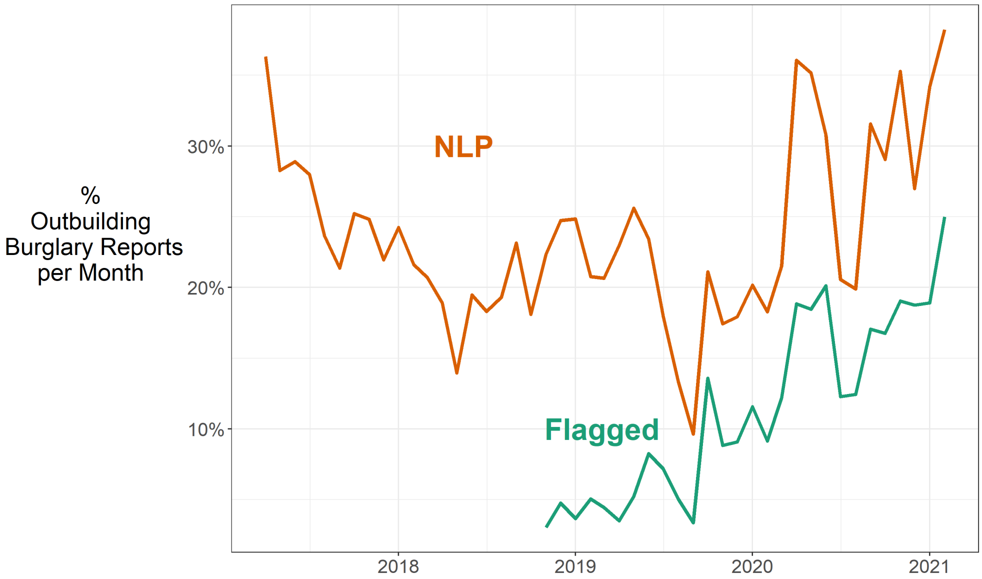


Figure 12.7: A time series plot of the outbuilding only classification. Showing data generated form the PTM model (NLP) and flags. Source: Author generated.

Test Set	Model	PF2 Model PF1 Data	PF1 Model PF1 Data
18/19	Motor vehicle	0.93	0.98
20/21	Motor vehicle	0.80	0.90
18/19	Force	0.91	0.93
20/21	Force	0.90	0.94

Table 12.5: MCC scores for the use of models built with PF1 data and used to classify PF2 data. PF2 metrics included for comparison.

12.4.5 Model transfer-ability

This subsection reports on the usage of models from one police-force area in another police-force area. The results are for the use of the PF1 models on the PF2 data; the reverse analysis could not be conducted for data security reasons. The MCC results in Table 12.5 show that the models are reasonably transferable. In each case, the MCC of the transferred model is lower, which accords with expectations. However, the drop is not particularly significant in all cases. This finding demonstrates that models that are built with data from one area can be useful in another area.

12.4.6 Performance over time

Even though the training data only cover the 2018-2019 period, the test sets were built for both 2018–2019 and 2020–2021 so as to enable observation of the variation in model performance over time. The results in Table 12.3 show the results from the 10 model initialisations in which the active learning data were used to fine-tune the PTM. The mean result for the 10 initialisations is reported here. There is a sizeable drop in the performance of the motor-vehicle model, from 0.98 to 0.90. This said, 0.9 may still be adequate, depending on usage. The performance of the use-of-force model improves slightly, from 0.93 to 0.94 (note that the highest-scoring run of the 2020-2021 set is superior to that of the 2018-2019 set). The performance of the outbuilding model also improves. The improvement is larger, with the relevant score increasing from 0.90 to 0.94.

12.5 Discussion

This section synthesises the results that were presented in the preceding one in the context of the main research questions that were described at the start of the chapter. Each question is explored in turn, and the results are compared to those from the original study, which draws on PF1 data.

12.5.1 Can the results in Study 1a be replicated in a different police force?

Study 1a set out to determine whether PTMs can be utilised to classify MO texts. The two classification tasks were 1) “Was a motor vehicle stolen during the burglary?” and 2) “Was force used to enter the building during the burglary?”. In addition, Study 1a inquired whether the PTMs are explainable and whether they work in the way that a human might do, that is, without relying on potentially spurious correlations in the data. In Study 1a, bias was examined to a limited extent due to a lack of data on victim characteristics.

The performance results in the replication study were equivalent to the results from the original study. Both resulted in high MCC scores, indicating that high-performing models can emerge from the fine-tuning of PTMs. An additional classification problem was explored in the replication study, namely that of outbuilding-only burglaries. A model was also fine-tuned for this problem, and it exhibits appropriate performance and a high MCC score.

If one compares the labels that were generated from the PTMs to the police-generated flags, one finds that the PTMs return a much higher number of crimes. Combined with the high MCC scores and the error analysis of the linked data, this finding suggests strongly that the PTM pattern in question is a more accurate reflection of intra-crime variation. Again, this result highlights the advantages of PTMs over existing police processes for exploring intra-crime variation.

Explainability was tested by having the LIME model generate word clouds which showed the most important words for each classification model. As with the PF1 models, the replication study produced word clouds that enhance trustworthiness. The important words that are highlighted in these clouds are entirely consistent with the words that a human may use to make a classificatory judgement. Therefore, they indicate that the model classifies similarly to a human.

More data on victim characteristics were available in the replication study than in the original study. Therefore, the models could be explored so as to detect bias against individuals of certain ethnicities and sexes. The results show that there is no evidence of systematic bias in the classifications of the fine-tuned PTMs. It should be noted that the text seldom made reference to the characteristics in question.

Therefore, bias was only likely to be introduced indirectly through systematic variation in the language and/or quality of the MO rather than through explicit references. It emerged from the bias investigation in the original study, Study 1a, that the length of the texts and the percentage of BERT words were not correlated with either of the bias metrics. In consequence, even if the MO texts on, say, Asian victims, had been short, the model would not have necessarily performed poorly in classifying them. There may be a number of different ways in which biases can be introduced into the chain that leads from a crime being committed to the formulation of an MO text and its subsequent classification. Firstly, the crime may not be recorded because the victim may prefer not to interact with the police; in such cases, there is no MO text. If the victim does interact with the police, the interaction might be suboptimal (e.g., due to language barriers). Consequently, the information that is available might not be sufficient for an accurate and comprehensive description of the crime. Finally, a PTM is built on data that are scraped from the Internet. These data are almost assured to reflect common biases in society and may perpetuate them through the classifications. The bias investigation in this study only concerns the last problem, that is, the use of the PTM. The first two channels by which bias is transmitted are beyond the scope of the study, as explained previously. The results here indicate that the biases that are inherent in the PTM do not affect the classification of burglary MO texts in the context of the particular classification tasks under observation.

In summary, this study replicated the satisfactory results from the original and extended them, proving that PTMs perform well when tasked with the classification of burglary MO texts. In addition, the models classify the texts by using words that are similar to the ones that humans would use, offering evidence in favour of the proposition that the models are trustworthy. The limited investigations revealed no evidence of systematic bias in the model classifications.

12.5.2 Can models trained in one police force area be used in another force?

Replicating the first study with data from a second police-force area highlights opportunities for transplantation. If model performance is unaffected by transposition, then the utility of the model is higher because models can be reused across forces without the need to share data. The results have shown that models can be transferred from one police force to another while retaining a reasonable level of performance. One implication is that forces can share models for direct use or seed the start of the fine-tuning of a separate model and therefore reduce the labelling burden. The practical implications may be significant, for example if the knowledge that is needed to classify a model is relatively specialised, as in the case of modern-day slavery crimes.

That models can be reused across forces also has implications for the implementation of PTMs. In the UK, for instance, there are 43 police forces, all with similar crime-recording techniques, a common language, and similar resource pressures. A central repository of models would be useful to all forces. Such a repository would allow sharing to be maximised and result in a commensurate reduction in the labelling burden. In addition, the technical aspects of model running and finetuning could also be conducted centrally, reducing the training burden across the 43 forces. Expanding the sharing of models to such an extent would require much more extensive experimentation than what this study, with its sample size of 2, can offer. Nevertheless, the results that were presented on the preceding pages are encouraging.

12.5.3 Are fine-tuned PTMs accurate over time?

As language use changes, so does the performance of models. The language of an MO text reflects intracrime variation. If that variation changes, for instance in response to the adoption of a new security technique, then so does the language of the MO texts. Models therefore have to be examined in order to ensure that they remain relevant to the language that is used. The models in the study were trained on data from one year then tested on data from a subsequent year. There was no perceptible drop in performance in either of the three classifications tasks. It appears that the models are robust to some changes that occur over time and even to significant disruptions such as the Covid-19 pandemic. However, despite the general decline in burglary, there is no evidence to suggest that a new type of intra-crime variation emerged. Variation that may have changed the language being used in the second time period used for this experiment. The evidence of the robustness of the models to the passage of time is limited, and there is no evidence of robustness to new criminal techniques and the resultant changes in language. Changes such as these need to be monitored, and the findings that were presented here certainly do not imply that the validity of finetuned PTMs does not need to be re-examined as time passes.

12.6 Conclusions

This replication study provided additional evidence for the proposition that PTMs can classify police MO texts effectively by extending the problems of the original to another police-force area and to an additional classification task. In addition, it was shown that there is no evidence of classificatory bias on the basis of either sex or ethnicity. The replication study also investigated the performance of models over time, finding no perceptible drop in classificatory power. This indicates the models will remain useful over extended periods of time. However, the study was relatively weak, and a more thorough study would be required for a definitive assessment of the rate at which models ought to be refreshed.

The models were also shown to be effective in solving the same problems with data from different police forces. This finding suggests that it may be possible for forces to share models. Model sharing would reduce the labelling, computational, and skills burden of using PTMs considerably. This may prove important in the practical implementation of PTMs because it would significantly reduce costs. It could also indicate that the centralised co-ordination and, perhaps, the development of some aspects of the relevant labour would be efficient.

These studies showed that PTMs can be effective when used with MO text data and across a number of different classification problems. However, MO texts are not the only texts that police forces generate. Police incident logs contain both crime and non-crime data. The next case study builds on this work by using PTMs to classify antisocial behaviour incident logs.

Chapter 13

Study 2: PF2 ASB Incident Logs

13.1 Introduction

The last study explored MO data and the use of PTMs to classify texts. This study investigates the applicability of PTMs to the classification of police incident logs. Police incident logs are text documents that are generally written by call handlers as they manage calls for service from the public. Incident logs are important because police forces do not only tackle crime. In fact, up to 90% of calls for service are not related to crime (Boulton, McManus, Metcalfe, Brian, & Dawson, 2017) and may therefore only be recorded as incident logs. By way of reminder, POP is also not limited to crime prevention; it is intended to reduce all types of harm for which the police can be deemed responsible. An investigation of the automatic analysis of incident data is therefore important because it can provide insights into a wide array of problems that police forces encounter.

This chapter explores the classification of antisocial behaviour (ASB) incident logs. These logs are a subset of incident logs that have been deemed to represent ASB. The remainder of this introduction briefly defines ASB before introducing a research article that influenced then applied the results of this study to investigate ASB during the pandemic in the UK.

13.1.1 ASB Definition

A recent briefing paper that was published by the House of Commons Library (Brown & Sturge, 2021) defined ASB as follows: "Anti-social behaviour (ASB) encompasses criminal and nuisance behaviour that causes distress to others. Typical examples include: noisy neighbours, vandalism, graffiti, public drunkenness, littering, fly tipping and street drug dealing." Legally, ASB has also been defined in two different contexts, namely the residential and the public. In both cases, the definitions are broad and revolve around the impact of the actions in question instead of defining them. Therefore, ASB is difficult to define precisely. In essence, it involves activities that have a negative impact on others but fall short of being crimes.

13.1.2 Published work

This study overlaps with the work that the author completed as part of the ESRC project Reducing the Crime Harms of the Covid-19 Pandemic. The author was part of a small team that published a related journal article (Halford, Dixon, & Farrell, 2022). It explored the effects of lockdowns on reports of ASB. The results from the present study were used directly in that article. The PTMs and the classification tasks that were used in this study were likewise directly influenced by the demands of that article. Therefore, the work behind this study was driven by two high-level questions. 1) Are PTMs useful for classifying police incident text? 2) How did ASB reports change during the Covid-19 pandemic? The first of these high-level questions reflects the true purpose of the study and forms the subject matter of this chapter. The second question facilitated the formulation of the question set for the first objective and is therefore explained in the next section in order to provide context.

I conducted all of the analysis that is presented in the journal article (Halford et al., 2022), of which the NLP work constituted approximately a third. In particular, I was the author of the data section, the methods section, and the NLP appendix.

13.1.3 Problem overview

The first Covid-19 cases in the UK were confirmed in January 2020, and they marked the start of the Covid-19 pandemic in the country. Shortly thereafter, in March 2020, the UK government imposed a national lockdown that restricted movement across the country and confined its citizens to their homes for long periods of time. Much has been published about the effects of the lockdowns on crime (see (Halford et al., 2020) for an initial review on an area in Northern England and (Langton, Dixon, & Farrell, 2021) for a longer-term perspective on the impacts in England and Wales). Against the general backdrop of a decline in crime during the pandemic, the police recorded a significant increase in reports of ASB. The increase in ASB was initially thought to be due to lockdown breaches being recorded as incidents of ASB.

The aim of the research paper was to investigate the cause of the increase in reports of ASB. In particular, the research question that animated the Covid-19 project was “Were reports of people breaching Covid-19 legislation the main cause of an increase in ASB reports?”. The alternative theory was that traditional forms of ASB, such as excessive noise, became more widespread in consequence of the de facto increase in population density.

Since ASB is not a crime, recording practices are not as rigorous as for crime data. Consequently, the ASB data that were available were unstructured and did not enable an examination of intra-incident variation or its variance during the lockdowns. An additional structured data field was introduced during the lockdowns, namely a ”Covid” marker that the call handlers could use if an incident was related to the coronavirus. However, the police analysts were not confident that this marker had been used consistently or comprehensively due to the speed with which it had been introduced. Therefore, NLP models were used to classify the data, and the changes in these classifications were observed over time. These classifications are explored in the next section.

13.1.4 Classification Tasks

Like the previous studies, this one focuses on the use of PTMs to classify police texts. The three classification tasks for this study were picked with a view to answering the questions that are related to the effects of the Covid-19 lockdowns on ASB. Those classification tasks are explained below. Examples that enable the classifications to be differentiated are given in Table 13.1.

1. Traditional ASB. The first question was whether an incident was a form of traditional ASB or not. ASB can encompass a wide variety of activities. Another formulation of this question runs as follows: "Could this ASB incident have happened before the pandemic?". If it was related only to a Covid-19 incident, then it could not have occurred before the pandemic had started. However, if it was a party or a noise complaint, then it could have happened before the pandemic as long as the complaint did not focus solely breaches of the Covid-19 regulations.

2. Covid Complaints. The second category is only related to the presence of a specific complaint about a breach of the Covid-19 regulations. Reports of failing to wear a face mask is a possible example.

3. Groups. This final category is also related to whether the ASB log contains a complaint about a group. Groups were assumed to comprise three or more individuals. References to families were excluded. For example, the text "Four adults having a party in a garden" would be assumed to refer to a group.

13.1.5 Article Conclusion

The conclusion of the article is that ASB reports did increase and that the increase was due in part to reports of Covid-19-realted infringements. Approximately half of these additional complaints also referred to a traditional ASB (e.g., noise complaints). Figure 13.1 is taken from (Halford et al., 2022) and provides a graphical summary of the results from the NLP analysis. The blue bars represent reports of traditional ASB, and the black line is a forecast of the ASB levels that would have been expected had there been no pandemic. They were generated through the use of a time-series forecast. The purple bars are ASB reports that refer both to traditional ASB and to Covid-19-related complaints (e.g., failure to wear a mask). The red bars are ASB incident logs that only refer to Covid-19-related matters. In general, the level of conventional ASB was consistent with expectations, and the additional reports included references to Covid-19 regulation breaches. One question that could not be

Example Text	Traditional ASB	Covid Complaint	Group
an email request has been made . default email notification has been made to xxxxx . com . email received xxxxx xxxxx 22/10/2020 22 xxxxx 12 incident relates to xxxxx group time of incident xxxxx 22 xxxxx 05 date of incident xxxxx additional information xxxxx i believe my neighbours are currently having a party with people outside of their household . i also believe that they have done this a few times recently . location address xxxxx flat xxxx , the village , xxxx xxxxx road , xxxxx xxxxx name of persons involved if known xxxxx is the subject displaying any covid 19 symptoms xxxxx unknown	N	Y	Y
- INFORMANT reporting there are 6 young men on motorbikes on the xxxxx way , riding round - INFORMANT said he cant see regs and DOESN'T want to get up close to them , - INFORMANT said they are right to the xxxxx way - xxxxx to covid-19 this is low asb and	Y	N	Y

Table 13.1: Examples of ASB incident logs and the labelled classifications

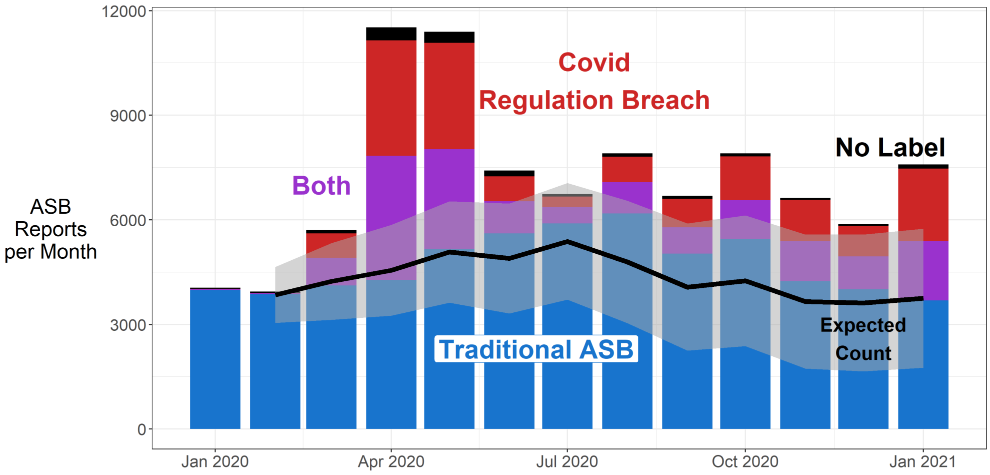


Figure 13.1: A plot showing recorded ASB for one northern police force during the covid-19 pandemic. Reproduced from Halford, Dixon, and Farrell (2022)

answered, however, was whether the increase in reports of traditional ASB was attributable to additional ASB or too the lower reporting threshold (the reporters had an additional reason to call the police, namely the Covid-19 infringement).

The remainder of this chapter follows the same format as the earlier studies that explored the utility of PTMs when they are applied to free text. The specific research questions for this chapter were set out above in the classification tasks Section. Next will be a review of the data, then the methods, the results, and, finally, a discussion and a conclusion.

13.2 Data

The data that were used for this study consist of ASB incident logs from PF2 for 2020. There are 93,809 logs in total. Incident logs were only included if their final classification was ASB. A detailed description of the ASB data was provided in Chapter 8. To reiterate, there are three main differences between the MO data that were used in the earlier studies and the incident log data that are analysed here. The first difference is of length – the incident logs are much longer than the MO texts. The median word count for the MO texts is 31, and the median word count for the incident logs is 166. Secondly, the police incident logs are also generated differently. They are ongoing logs of the events that transpire in the course of an incident. The logs are rarely edited; instead, they are generated by operators in control rooms as incidents unfold. Incident logs are intended only for internal use, whereas MO texts are generally written post hoc by police officers and intended for external use. Accordingly, the names of suspects and other forms of personal information are not used routinely. Thirdly, although the same whitelisting procedure was applied to texts of both types, that process was tailored to the redaction of MO data and not to the redaction of ASB logs. Coupled with the different generation process, this feature of the problem means that a higher proportion of words were redacted in the ASB logs (8%) than in the MO texts (2%). Close to one in every 12 words in the ASB logs was redacted.

13.3 Method

13.3.1 Data Labelling

The data were labelled by two researchers according to the classifications that were outlined earlier. Disagreements between the two labellers were settled by the author. The data were selected by using active learning and on the basis of the Covid-19 complaint classification task. As before, a test and validation set were randomly selected before the training set was developed. The batch size for active learning was set to 50. Again, this is roughly equal to 1 hour of labelling for each batch of texts. A total of 900 incident logs were labelled, with 200 logs labelled for both the validation and the test sets and an additional 500 logs labelled for the training set. Labelling ceased when the resources of the researcher had been expended.

13.3.2 Fine-tuning the PTM

Since the incident texts are generally longer than the MO texts, it was not possible to use the BERT model from the previous studies. As mentioned in Chapter 9, the Longformer PTM (Beltagy, Peters, & Cohan, 2020b), a similar model, was designed for longer texts. The Longformer model was used throughout the study for the classification of police incident logs. The length of the text still posed problems for the computing facilities that were available, particularly computer memory. For this reason, the hyperparameters of the model were adjusted to avoid memory problems (i.e., attempting to use more memory than was available) rather than optimised for model accuracy. Even after the adjustment of the hyperparameters, the maximum text length had to be set to 1,500, meaning that the final words of some (< 1%) of the incident logs would have to be removed when the logs were entered into the model.

In addition, during the fine-tuning of the model, it was discovered that removing the “xxxxx”token from the incident logs improved classification performance. In the whitelisting process, the “xxxxx” token replaces words that are not on the safe list, typically nouns. Therefore, all model fine-tuning was conducted with the “xxxxx” token removed from the logs.

13.3.3 Performance

Like in the previous studies, performance is measured by reference to MCC metrics in order to determine the accuracy of the model. Explainability is explored through the use of the LIME tool. Word clouds were generated. They contain the most important words for each classification. No metadata on victim or offender characteristics can be extracted from the police incident logs. Therefore, the investigations of bias are limited once more. However, as will be shown later, the investigation of explainability indicates that the word “Default” may have had an undue influence on the model classifications. On further inspection, it emerged that the word “Default” is used when an incident log is generated from a complaint that is submitted via electronic means, that is, when a member of the public files their complaint through the online system or via email. For this reason, the partition for the bias investigation is based on request method.

The term “request method”refers to the channel by which a request is received. Typically, requests are received by phone or electronically (through online forms or by email). The data were split into three categories, namely “telephone” (including emergency and nonemergency calls), “electronic” (including online forms and email), and “other”(including logs generated by officers). Since the bias investigation is limited to binary splits of the data, two partitions were required for each classification. The two partitions are 1) “telephone” versus “electronic” and “other”, and 2) “electronic” versus “telephone” and “other”.

These two partitions were examined for each classification type. Therefore, there are six bias values for each metric. As before, the test set was investigated and a separate tenfold cross-validation experiment was used to understand the potential range of values from 10 different models that were trained on different and randomly selected data, as explained in Chapter 8.

13.4 Results

13.4.1 MCC

The MCC metrics for the ASB police incident logs are generally lower than in the earlier studies. No classification model achieved an MCC of more than 0.9. As in the earlier studies, each model was built 10 times in order to explore variation due to randomness. There was considerable variation across model builds. Variation occurs due to the random initialisation of the models. The Groups classification of police incident logs that had the highest MCC score (0.83). It was followed by the Covid classification (0.81) and then the Traditional ASB (0.71) classification. The F1 scores were recorded for comparison. The F1 scores are comparable to but lower ($\approx$ 0.05) than the scores from the Longformer models that were fine-tuned on standard academic NLP tests (see Table 7 of Beltagy et al. (2020b)).

13.4.2 Explainability

As in the previous studies, the results for explainability are presented as word clouds. A satisfactory result is recorded if the larger words within the cloud have an intuitive bearing on the classification type. Unlike in previous studies, there is only one word cloud for each classification type. Producing word clouds for the police incident data required a temporary increase in computer memory. However, the idea of producing both negative and positive word clouds for each classification was only implemented after that temporary increase had ended. Accordingly, only the positive word clouds are available.

The word cloud for the traditional ASB classification is displayed in Figure 13.2. Compared to previous word clouds and other ASB word clouds, there are

Run	Trad ASB	Covid	Groups
1	0.59	0	0.78
2	0.63	0.78	0.79
3	0.67	0.62	0.8
4	0.68	0.73	0.8
5	0.66	0.81	0.81
6	0.52	0.81	0.77
7	0.59	0.75	0.83
8	0.64	0	0.81
9	0.71	0.74	0.79
10	0.67	0.73	0.77
Mean	0.636	0.597	0.795
Best Run	0.71	0.81	0.83

Table 13.2: Table of ASB model metrics MCC scores for the three ASB classification problems. Each model was trained 10 times with the same data.

not a few select words that influence the prediction. The word cloud contains many words of a similar size, indicating that the words in question have similar impacts on classification. This was analogous to the word clouds in Study 1, in which there was no direct mention of the nonoccurrence of events (such as cars not being stolen; see Figures 10.2b and 12.2b).

Figure 13.3 displays the word cloud for the classification of Covid complaints. This word cloud contains the most important words for determining whether an ASB log contains a Covid-19-related complaint. The largest word is “Covid”. This is not surprising, and it indicates that the model is working as expected. Another notable word is “Default”. There is no obvious connection between a Covid-19-related complaint and the word “Default”. Though as explained earlier it is revealed how the model uses that word.

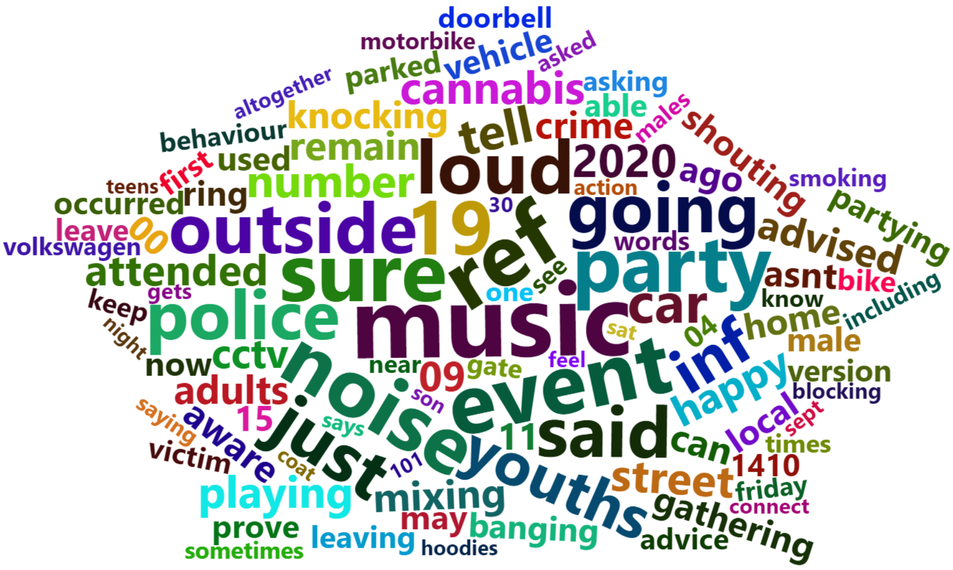


Figure 13.2: Word cloud for traditional ASB classification. The top 100 words that contributed to a positive classification of traditional ASB. Source: Author generated.

The third word cloud is related to the Groups classification. This word cloud is displayed in Figure 13.3. The largest words are "group", "party", and "groups". These words are clearly related to groups or gatherings and indicate that the model is working on words as expected. another significant though less prominent word is "males". This is possibly prominent because most gatherings that are related to ASB are primarily or exclusively attended by males. However, if the model seeks male groups exclusively, female groups may be more difficult to identify. In other words, there may be a bias towards males being identified as members of groups. Unfortunately, without data on the sex of the (potential) offenders, this potential bias cannot be explored systematically here.

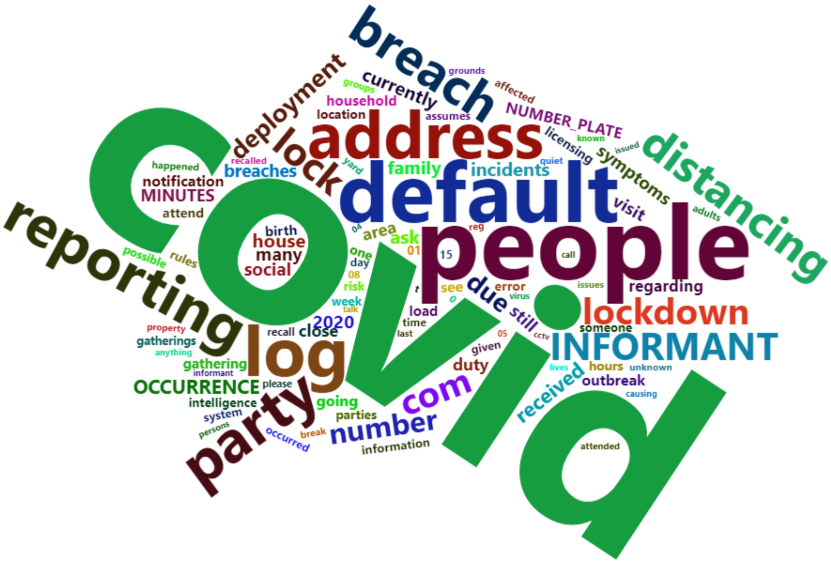


Figure 13.3: Word cloud for Covid ASB classificationWords that contributed to a positive classification. Source: Author generated.

13.4.3 Bias

Table 13.3 displays the results from the bias investigation. The bias investigation focuses on complaint transmission methods (by telephone or electronically). This investigation revealed evidence of bias. The results are explored by classification type. In each case, the electronic partition is approximately the negative value of the telephone partition. This finding is not unexpected because the "other" category is smaller than these two categories. It was included in order to ensure that it would not have a strong effect, which it appears not to have had. The result is that the results can be described exclusively by reference to the "electronic" partition.

The first classification type is "traditional ASB". The p values for the tenfold cross-validation experiment indicate that there is bias at a statistically significant level of both EoO and PP. However, there is a disparity in the sizes of the bias between the value of the metric from the test set and the mean value from the CV set. For EoO, the bias is larger in the test set; for PP, the bias is larger in the CV set. For EoO, which has a negative value

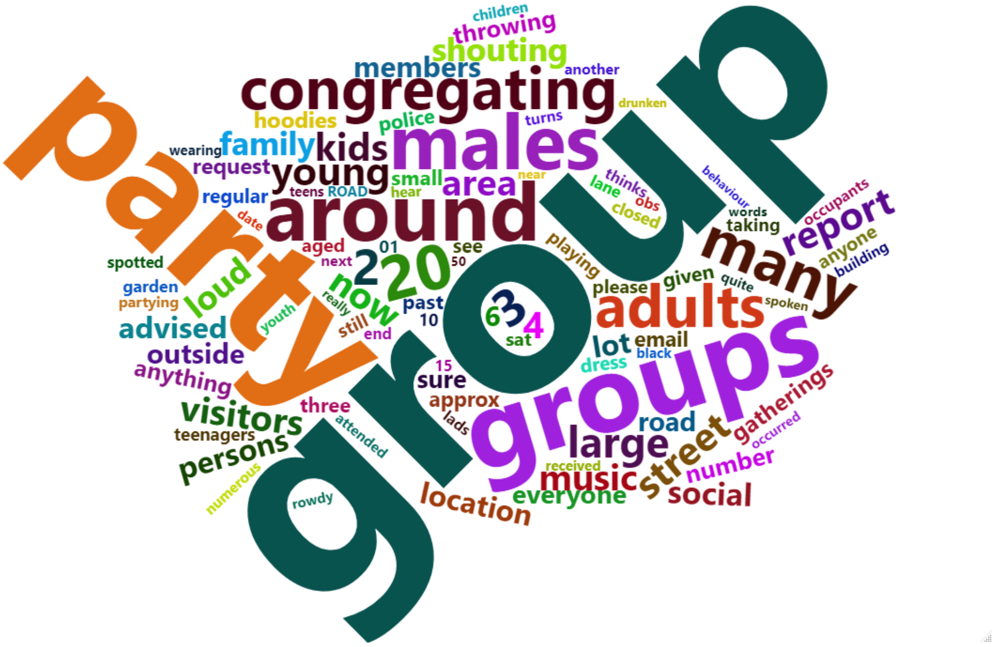


Figure 13.4: Wordcloud for group ASB classification. Words that contributed to a positive classification. Source: Author generated.

for the electronic partition, the foregoing supplies evidence of bias against requests that are submitted by electronic means. In other words, the recall power for the electronic methods of making requests is lower than that of the telephone methods. The PTM finds it harder to classify positive instances of traditional ASB reports that are submitted electronically than to classify telephone reports. The values for PP are also negative, indicating that the electronic requests are subject to more errors. As far as PP is concerned, the equivalent result, formulated in words, would be as follows: among the positive instances that the PTM found, the accuracy of the positive classifications was lower if a request had been made electronically.

The next classification type is Covid complaint. The results for EoO and PP are similar. The test set produces larger absolute-sized metrics than the CV set. All CV metrics are statistically significant. However, unlike for the Traditional ASB classification, the signs of the metrics are reversed. In this case, the bias is against the telephone request method. The size of the biases for the Covid complaints are larger than the equivalent metrics for the traditional ASB classification. The EoO outcome is as follows: the PTM finds it easier to classify positive instances of Covid complaints correctly when they are submitted electronically rather than by telephone. The PP outcome, in words, is as follows: among the positive instances of Covid-19 complaints that the PTM found, the accuracy of the positive classifications was higher if the request had been made electronically.

The final classification type was the Groups classification. The group classification has the smallest bias metrics. Unlike for the other two classifications, the directions of the bias are not consistent, and not all of the metrics are statistically significant. The evidence of bias is therefore weaker for this classification than for the other two. The EoO has mixed signs across the test set and the CV set results. The CV set results are statistically significant. The values of the metrics are the smallest across all three classifications; if there is bias, its effect is small. The evidence of PP bias is weaker still, with the CV mean no longer statistically significant. The conclusion is that there is no strong evidence of EoO or PP bias in the group classification.

13.5 Discussion

This section discusses the results that were presented in the preceding one, particularly by reference to the previous studies that explored the MO data. The findings of the two sets of studies (MO and Logs) are consistent. However, there are some important differences that are explored on the pages that follow.

13.5.1 Performance

The MCC metrics for the incident data were lower than the metrics for the MO data. The models classified incident texts less successfully than MO texts. This finding is likely attributable to three causes. Firstly, the incident texts were not edited, and so were contradictory in places. Reading and comprehending such texts is often difficult for humans. Secondly, less training data were available. Only 500 texts were used to train the models for the incident texts, whereas more data (700 and 900 texts) were used to train the models in Study 1a. Thirdly, the model architecture was different. Recall that the Longformer model was preferred over the BERT model for the incident texts. The next paragraphs explore each of these matters in turn and then inquires whether

Equality Of Outcome				
Model	Partition	Test Set	CV Mean	CV p value
Traditional ASB	Electronic	-0.128	-0.067	0.005*
Traditional ASB	Telephone	0.103	0.060	0.003*
Covid complaint	Electronic	0.264	0.190	0.002*
Covid Complaint	Telephone	-0.264	-0.175	0.002*
Group	Electronic	-0.023	0.039	0.007*
Group	Telephone	0.018	-0.037	0.006*
Predictive Parity				
Model	Partition	Test Set	CV Mean	CV p value
Traditional ASB	Electronic	-0.007	-0.151	0.002*
Traditional ASB	Telephone	0.000	0.139	0.002*
Covid complaint	Electronic	0.435	0.166	0.003*
Covid Complaint	Telephone	-0.375	-0.160	0.008*
Group	Electronic	0.049	0.014	0.610
Group	Telephone	-0.004	-0.009	0.754

Table 13.3: Extrinsic bias metrics for the PF2 ASB models. The model is denoted by the classification task. The partition is the factor used to split the data. The test set metric is calculated from the original test set. CV refers to cross validation. *CV mean* is the mean of metrics from the 10 fold cross validation experiment. *CV p value* is the p value for the hypothesis that the CV mean value is not zero. * indicates a p value that is significant.

the models are suitable for use.

The data

The incident texts were not edited documents. They contain boilerplate text, and they were redacted. All of these factors mean that the data to be predicted (the incident texts) were dissimilar to the original training data on which the PTMs were first trained (the pretraining data). When the data to be predicted are different from the pretraining data, models become less powerful because they do not learn the representations of that language well.

In broad terms, there are two solutions to this problem. Firstly, the model can be pretrained on similar data from the outset. In this case, incident log data can be used much earlier than Wikipedia data. However, such a solution would require vast amounts of data and computing power.

Secondly, if the model can not be changed then perhaps the data can. Can the incident texts be changed so that they become more like the training data?If so, how? Words can be added to the model dictionaries to enable the model to represent more words. Jargon within the incident texts can be translated into more widely understood words. The data cannot be redacted, meaning that security needs to be met in other ways. Boilerplate text can be removed. Every transformation of the data would require additional effort. If the quantities of data are large, then the transformations would need to be automated, limiting the scope of the changes. Which solution works best – changing the data or changing the model? This remains an open question that ought to be tackled in future research.

Labelling

Fewer data were labelled for these classifications than in the earlier studies, which may have resulted in a lower MCC score. In addition, the larger variation in MCC score across the 10 random initialisations also indicates that the model did not converge on the optimal solution. More data were required because the of the larger variation within the texts. Therefore, future researchers, when confronted with longer unedited texts, may wish to allocate more resources too labelling.

Computing power

Although the computing power that was available for the modelling task was adequate, in that the models could be run, it was suboptimal because the hyperparameters had to be adapted to in order ensure that the amount of memory required for modelling would be lower. Unfortunately, this limitation is a factor when PTMs are used with long texts. Machines with access to larger amounts of memory do exist, but they are not typically desktop computers. Longer texts are problematic because such computing facilities are not routinely available to police forces. This said, the development of cloud technology may make access to more powerful computers less difficult over time.

When is a model good enough?

"All Models are wrong, but some are useful[1] ".When is a model good enough for use? In other words, what MCC score needs to be achieved for a model to be useful? This is an open question that has no definitive answer, but a decision can be made by considering three questions. Firstly, at what scale is the model intended to be used? In this instance, the models were used to track the change in ASB over time. Relative rather than absolute changes were observed, and the correctness of single instances was not excessively important relative to overall consistency. However, if the response to an individual instance does matter, then a higher MCC score is clearly better. Secondly, how good is the model when compared to an existing process? Typically, the existing process is that of humans reading texts. Humans are not infallible. They suffer from fatigue, and they are generally expensive. A total of 93,000 texts were classified. As a conservative estimate, a single individual would need 124 working days, that is, 24 working weeks (or approximately half of the working year), to read those texts. In short, without the PTM, the work would not have been completed. Thirdly and lastly, what is the cost of an error? There are two possible ways to commit an error, a false positive and a false negative. The costs of the two may differ. For instance, the cost of missing a domestic abuse crime may be larger than the cost of expending resources on a crime that is not domestic abuse.

[1] Attributed to George E.P. Box.

Each problem entails specific cost-related trade-offs, and these trade-offs affect the robustness requirements for the model.

In short, although the performance metrics that were used here can point to models that are superior to others and quantify the likely errors, they cannot be used in isolation to determine whether a model should be used. Important considerations also emerge from the investigations of explainability and bias that follow.

13.5.2 Explainability

The explainability results for the incident text models are similar to those for the MO models in Study 1. Specific mentions of a classification, such as a Covid-19 complaint, were associated with related words that were more prominent. The traditional ASB classification, which is not related to a specific type of incident, produced a more homogenous word cloud, reflecting the wider spread of possible descriptions.

One notable exception emerged from the explainability investigation. It was highlighted by the word clouds. This exception was the prominence of the word “Default” in the Covid classification. The word “Default” is used primarily when an email or an online complaint is added automatically to the police incident logs. The word thus denotes the channel by which the complaint was received. On the whole, 64% of online and email reports contain Covid-19 complaints. The corresponding figure for the telephone reports is 22% (see Figure 13.5 for all percentages). Therefore, a complaint made by email or online was much more likely to have been a Covid-19 complaint than one made over the telephone. This can be seen further in the bias statistics. For the Covid classification, the misclassification rates differ between the electronic and the telephone delivery methods.

This is a good example of of the importance of explainability investigations and of their usefulness for improving predictive accuracy. In this case, removing the standard text that is added to the electronic forms of incident logs is likely to improve classification because the model cannot use a proxy for logging type. This text was not removed here, which opens an avenue for future research.

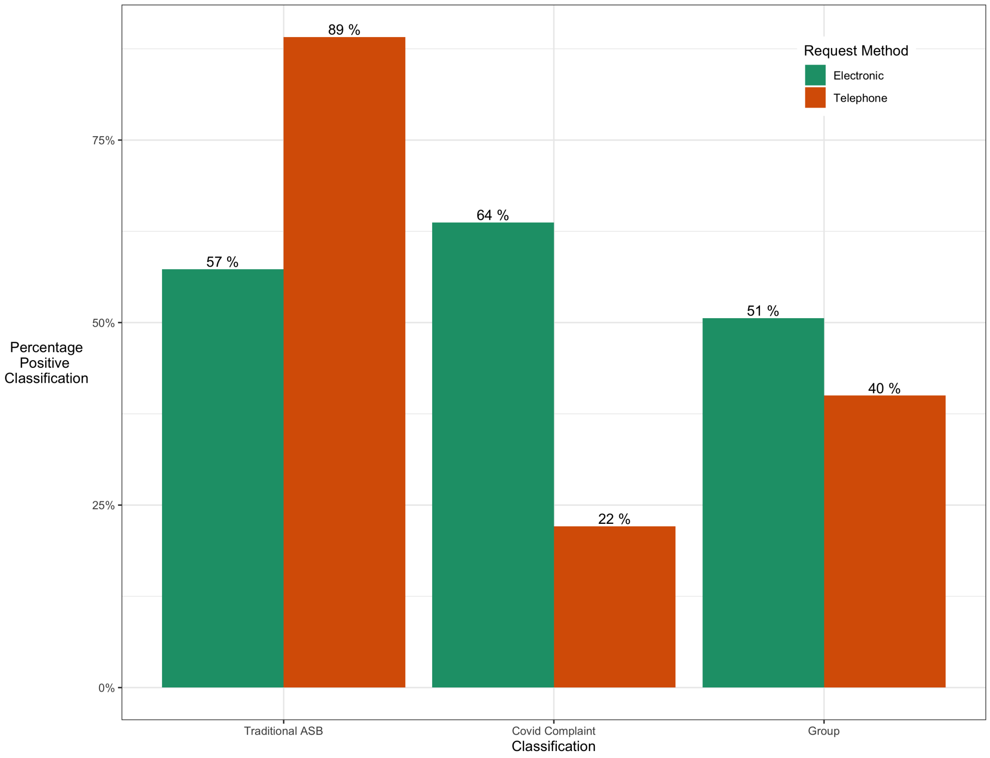


Figure 13.5: ASB positive classifications by request type. The proportion of each positive classification by request type. For example 89% of telephone requests contained traditional ASB. Percentages calculated from all of the hand labelled data. Source: Author generated

13.5.3 Bias

Although the data could not be examined for bias against victims or offenders with particular characteristics, it could be examined for bias against particular request methods. The results contain evidence of bias for two of the classification types. For traditional ASB, the PTMs were less accurate when classifying electronic requests. For Covid-19 complaints, the PTMs were less accurate when classifying telephone requests.

As shown in the explainability section, it was found that the Covid-19 classification method drew on the elements of the automatically generated text to make predictions about Covid-19-related complaints. This was reinforced with the bias metrics – the Covid-19 classifier made more mistakes when applied to the electronic data because it relied on automatically generated text (e.g., “Default”) that does not contain information about Covid-19 complaints but instead reflects correlations.

The traditional ASB classifier made more errors when applied to the telephone data. It is evident from Figure 13.5 that the telephone data predominantly contain logs that are related to traditional ASB. On this occasion, the explainability section supplies no evidence of words being misused. Therefore, it is not clear whether this bias is based on the same mechanism, that is, it is not clear whether there are words or phrases that distinguish the telephone data from the electronic data. A more profound investigation of the error analysis of individual logs and the word clouds for the negative classifications is necessary to understand the mechanisms of bias more adequately.

13.6 Conclusion

In conclusion, the results from using PTMs with police incident texts are encouraging, even if they are less satisfactory than the results for MO text. The inferior results are unsurprising. Before modelling, the length of the unedited texts and the larger loss from whitelisting were noted as factors that could contribute to inferior outcomes. These factors, coupled with the suboptimal computing power and the lower volumes of training data that were available, contributed to the lower MCC scores. Some of these issues may be overcome in future research, enabling the power of PTMs to be harnessed more effectively. Whitelisting can be tailored and perhaps even eradicated if data are kept on police servers. Text data can be modified to remove automatically generated text. More computing power can be resourced through the adoption of more adequate research plans. However, absent substantial developments in transformer model architecture, the length of the incident logs may continue to be problematic. This issue has implications for other long texts, such as witness statements, that police forces may wish to analyse. The next section summarises the whole chapter.

13.7 Summary

This chapter introduced and tested the use of PTMs with police incident data. The main difference between this study and the previous ones lies in the type of data that were analysed and the model that was used.

This study examined police incident data, which is both longer and less structured than the MO data that were used previously. More words were removed from the incident data through whitelisting, and the data constitute unedited logs that were written as situations were developing. The incident data also include stock phrases from online and email reports. The incident logs are generally longer than the MO texts.

The length of the data meant that the BERT model from the previous studies was unsuitable to the present ends. Therefore, the Longformer model was used. It is designed for longer texts. The length of the texts also meant that fine-tuning placed more significant demands on the available computer memory. Due to the limited memory at the disposal of the researcher, the hyperparameters were set so as to lower memory requirements rather than to optimise performance.

Three classification tasks were completed. These tasks were developed in order to answer questions about breaches of the Covid-19 legislation during the 2020 lockdowns. The classifications revolved around the presence of traditional ASB, around Covid-19 complaints, and around mentions of groups of individuals in incident logs. The data were labelled by using active learning, and the Longformer model was fine-tuned on these tasks. The performance metrics were lower than for the MO data. They were nonetheless comparable to standard benchmark tests. This is partly due to the scarcity of computing power and labelled data for model tuning.

The word clouds from the explainability investigation pointed to an anomalous word for the Covid-19 classifications, namely “Default”. Further investigation revealed that the word in question is predominantly used in incident logs that are based on electronic requests. The PTM used “Default” as a proxy for online reports, which contained a much higher proportion of Covid-19 complaints than others. This tendency generated bias against certain reporting channels in the classification of the incident texts. Victim bias could not be analysed due to a lack of metadata.

This study showed that PTMs can be used to analyse police incident logs at a large scale. More powerful computers would be needed to explore the full potential of PTMs. Biases will also have to be mitigated, especially if the data contain phrases that can be used as proxies for external variables.

This was the final case study in this thesis. The next chapter summarises the results from all of the studies that were conducted.

Chapter 14

Summary of Study Results.

14.1 Introduction

This chapter concludes the second part of this thesis. Its aim is to draw the lessons from the previous studies together by reference to the supporting objectives that were set out in Part 1. The research question is as follows:

Can PTMs be used efficiently to extract information from police free-text data, and if so what practical applications for problem-oriented policing does this approach have?

The supporting objectives are

1. **Identify the extent of NLP usage with police data.**
2. **Evaluate how effective PTMs are with MO data.** .
3. **Evaluate how effective PTMs are with Police Incident data.**
4. **Evaluate how effective Active Learning is with police data.**
5. **Identify which parts of the POP process might be best supported by the use of PTMs.**
6. **Identify implementation barriers for PTMs.**

14.2 Summary of Study Results.

This section refers to the supporting objectives to explore the cross-cutting issues that were identified in the studies, as outlined in the preceding parts of this thesis.

14.2.1 Extent of NLP usage with police data.

This was conducted in the literature survey in Chapter 6. The main findings from that chapter was that, despite the sporadic use of NLP models to analysis police-generated data in research, PTMs had not been studied previously. The study of the PTMs that are applied to police data is important because PTMs are currently the most powerful NLP models, as judged by academic NLP benchmarks, and their performance may therefore be superior to that of the methods that are currently employed to treat police data. In addition, Chapter 6 also indicates that, when they use algorithms and NLP models, the police are concerned about factors other than performance, such as bias within the models and explainability. These factors were examined in the studies and in the performance assessments, which also refer to bias and explainability as well as to the performance metric MCC.

14.2.2 Evaluate how effective PTMs are with MO data.

The evaluation of the application of PTMs to MO data was the subject of Study 1a and Study 1c. Study 1a investigated two different classifications by reference to PF1 MO data. Study 1c investigated three different classification tasks by reference to PF2 MO data. In addition, Study 1a compared the PTM models to a simple keyword approach, and Study 1c compared PTMs to an existing flag method. Study 1c also investigated the application of PTMs across police forces and over time. The results from the classifications of the two studies are discussed first, which is followed by an examination of the minor experimentations.

In Study 1a, two classification tasks were undertaken with burglary MO data. The first task entailed determining whether motor vehicles had been stolen during burglaries. The second classification task was to determine whether force had been used to enter buildings during burglaries. In both classifications, the fine-tuned PTMs exhibited high performance, with MCC scores above 0.97. In the motor-vehicle model, one fine-tuned model labelled every MO text in the test set correctly. The downside to these PTMs has to do with the labelled data that are required. PTMs are supervised learning models, which means that they require labelled data from which to learn. Approximately 900 labelled examples were needed, equal to approximately 9 hours, or 1.5 days, of labour. In the case of the PF1 data, there were 9,961 burglaries. Labelling 900 examples by hand takes less time than reading 9,961 MO texts. However, if the amount of burglaries of interest is smaller, for example because the area that is subject to a POP intervention is not large, then spending time on labelling data and fine-tuning a PTM may be inefficient.

Both studies also investigated the explainability of the models. Explainability is important because it is conducive to the formation of trust in the models and because police officers might need to explain how results are generated to interested parties, including the public. Explainability was investigated though the LIME tool. The investigation culminated in word clouds that highlight the most important words for each classification. The word clouds that were generated from the MO data contained words that a human might use to complete the same tasks. For example, the word “smash” is prominent in the use-of-force world cloud, indicating, as one would expect, that if the word “smash” is used, then a property was broken into. However, when an omission, such as not stealing a car, is not described, the word cloud can be inconclusive. An inconclusive word cloud is one in which the words are of a similar size and in which no words are prominent. This, however, is generally a reflection of the structure of the MO data.

Bias was also investigated in both studies. The bias examinations used two metrics, namely 1) EoO, which measures the disparity of the probability of TPs across groups, and 2) PP, which measures disparity of the probability of FPs across groups. These metrics were calculated for the test set and then for 10 cross-validation test sets.

Due to the differences in data availability across police forces, bias was investigated by using different factors for PF1 and PF2. For PF1, bias was investigated by reference to the statistical characteristics of the MO texts, namely 1) length and 2) the number of word pieces. Word pieces are important because they are a measure of how many out-of-vocabulary words are used in an MO text. An out-of-vocabulary word is a word that is not included in its entirety into the predetermined vocabulary of the PTM. For this reason, it is broken down into multiple pieces that are accessible to the model. For example, PTM vocabularies do not include the word “untidy” in their vocabulary. It is broken down into “un” and “tidy”. The bias investigation drew on the Pearson correlation coefficient to determine whether there was a relationship between the accuracy of model predictions and statistical properties. No evidence of a relationship between PTM performance and the number of word pieces or the length of an MO text was found – there was no evidence of bias. For PF2, the bias investigation was conducted by using victim characteristics, namely gender and ethnicity. There was no strong or consistent evidence of bias on the basis of either characteristic.

In addition, Study 1c investigated the use of PTM models over time and across police forces. There was no significant drop in performance over time, and PTMs that are fine-tuned for one force can be used by another. However, performance deteriorates across forces.

14.2.3 Evaluate how effective PTMs are with Police Incident data.

Police incident data are different from MO data. The main differences are that police incident data are longer and less well edited than MO data. The incident data also contain more words that are redacted in the course of the whitelisting process. Each of these differences may contribute to inferior PTM performance. Importantly, due to the length of the incident texts, a different PTM, Longformer, was used for the analysis.

The police incident data were explored in a similar manner to the MO data but only for one police force, PF2. Three classification tasks were used to explore the use of PTMs. These classification tasks were analysed in terms of performance (by using MC), explainability (by using LIME and word clouds), and bias (by using the bias metrics EoO and PP). The bias investigation focused on the method by which complaints are received (electronically or by telephone).

The performance of the PTMs when applied to police incident data was not as strong as their performance on the MO data. This said, the results from the police incident data were comparable to the performance of PTMs when applied to recognised benchmark datasets from the literature. Some, but certainly not all, of this decrease in performance may be attributed to factors that are specific to this study. These factors include 1) the need to redact text and 2) suboptimal computing power. Data redaction entails some loss of information. The redaction rate of the police incident texts (8%) was higher than that of the MO texts (2%). However, the police would not need to make such redactions. Therefore, if the PTMs had been used by the police, then performance would not have been hampered by redaction. Secondly, the computing power that was available for this study was suboptimal. Consequently, the PTMs could not be fine-tuned optimally, and maximum performance may not have been attained.

The investigation into explainability revealed that the PTMs were using system-generated text to aid prediction. System-generated text is text that is generated by the police computer systems as they process complaints that are submitted electronically, for example by email. This reliance on system-generated text was reflected in the investigation of bias.

Since there were no data on victim or offender characteristics, the bias investigation focused on the channel through which the reports were received. In broad terms, there are two methods – the public can make telephone calls (emergency and nonemergency numbers) or rely on electronic means (email or online forms). The investigation revealed that the PTMs were biased when classifying incident texts. The bias was consistent with the underlying base rates of the classification types in the delivery method types. For example, since a high percentage of Covid-19 complaints had been made electronically, the PTM was more likely to over classify logs as Covid-19 complaints if they had been delivered by such means. It is likely that this bias was due, in part, to the system-generated text, which effectively identified a given log as having been received electronically.

14.2.4 Evaluate how effective Active Learning is with police data.

Active learning is a method that is designed to reduce the overall volume of data that require labelling. Active learning achieves this objective through the incremental fine-tuning of a PTM and the use of the resultant model to select the next set of data to be labelled. The results from Study 1a show that there was, on average, a reduction of approximately 14% in the volume of data that had to be labelled. However, this efficiency was partially offset by the additional processing time that was needed to fine-tune the PTM in each round of active learning. Therefore, the results are not conclusive, and the desirability of using active learning depends, in part, on the time that an analyst has at their disposal (if time is short, active learning is useful because there are fewer data to label) and the relevant deadline (if close, then active learning with PTMs may take too long due to the additional processing time).

14.3 Study Limitations

This section reviews the limitations of the studies that were presented in this part of the thesis. The limitations are reviewed by reference to the aims of the study.

14.3.1 Problems

In general, the results from the studies are encouraging. However, the problems that were identified and used to guide the construction of the models are highly limiting. Only one type of crime, burglary, and one type of incident, ASB, were examined. Each of these problem areas only has three factors, yielding six different problem-incident combinations in total from what could be an infinite combinatorial space. In short, the sample is small. The PTMs were only proven to be useful for a small set of problems.

14.3.2 Data Types

Similarly, the diversity and the volume of the data were also limited. Only MO texts and incident logs were considered. Police forces have more document types in their data stores, including crime summaries and witness statements. Some of these documents can be long. The documents that were examined here are relatively short. PTMs were only shown to be useful when applied to data in which short documents predominate. Further investigations are required to determine whether this performance can be repeated with documents of different types, particularly longer ones.

14.3.3 Explainability.

Although explainability tools were employed, the results were not trialled robustly. Explanations are context and audience specific, and the explanations that were generated here were not trialled with those who might use them. In addition, the explanations that were generated rely on local models; the global effects of words are not fully understood. However, this deficiency was mitigated by the use of the LIME tool across 200 texts and the aggregation of the findings.

14.3.4 Bias.

Three important areas were specified for bias in Chapter 8. The three are data coverage, data selection bias, and algorithmic bias. Each type of bias can have an impact on the final results. The bias investigations in this study focus on algorithmic bias. The other two types of bias were beyond the scope of the research, but could have affected the results. In the case of POP, prevention resources can be allocated inequitably. Bias against victims with certain characteristics could only be investigated by reference to the PF2 MO data. This area would certainly need to be explored more thoroughly for individual models to be used operationally.

14.3.5 Conclusion

In conclusion, the use of PTMs with police data was successful. Performance was proven to be satisfactory, especially with MO texts. Limited evidence of bias was found; importantly, none of it revealed bias against victims with certain characteristics. In addition, the explainability tools showed that, for the most part, the PTMs complete classification tasks by using words that would make sense to a human reader.

The conclusion from these studies is that PTMs are useful for classifying police data. How might this classification be used for POP? Where in the SARA cycle could it have the strongest impact? These questions are answered in the next chapter, which applies the results and examines their limitations in the context of POP and the SARA framework.

Part III

Discussion

Chapter 15

Implications for POP

15.1 Introduction

This is the first chapter of the third and final part of the thesis. The aim of this part of the thesis is to draw together the lessons from the previous two parts and to explain their meaning for the future of POP and NLP. This chapter synthesises the results from the previous studies in order to meet Supporting Objective 5, “Identify which parts of the POP process might be best supported by the use of PTMs.” The next and final chapter focuses on avenues for future research on the use of NLP with police data.

This chapter has three sections. The first section refers to the SARA framework for POP and identifies areas of the framework where PTMs may be useful. The second section concerns two additional matters that are related to the use of PTMs, namely 1) the implementation style of POP in police forces and 2) the sharing of fine-tuned PTMs. The concluding section overviews the technical and physical barriers to the implementation of PTMs in the POP cycle.

15.2 POP Applications

Part 1 explored POP. A POP framework, SARA, was introduced as part of that exploration (see Figure 3.2). SARA is a four point framework, the elements of the framework are - Scanning, Analysis, Response and Assessment. Although there is a natural order to the framework, it should be stressed that moving back and forth is encouraged because it enables the process to be refined. The following sections briefly restates the aims of the elements of the SARA framework and explores the potential applicability of PTMs.

15.2.1 Scan for Problems

Scanning is the first stage of the process, and it revolves around finding and defining the problem that is to be solved. By way of reminder, a problem is a cluster of similar and related incidents that cause harm to the public and can be considered to be a police responsibility. Problems are not necessarily crimes. In fact, the ASB from Study 2 is an example of a serious non-crime problem.

Generally, scans are conducted on the basis of prior information, that is, the individual who scans already has an idea about, for instance, the type of incident and the variation that they are looking to identify. In this instance, an attempt is made to confirm or rule out that variation and to find other problems with the same characteristics. Alternatively, the scan may focus on problems of unknown form and variation, such as high-harm or novel problems. This second type of scan is examined at the end of this section.

The scan is conducted with a general idea in mind about the problem type. As an example we can use the second problem identified in Chapter 10 (10.1.1). The problem was to determine whether an outbuilding or a home had been burgled in each recorded burglary. The detective was aware of a spate of burglaries but believed that it had been the result of a high proportion of outbuilding-only burglaries. In this instance, the suitability of PTMs can be determined by answering two questions. 1) Are the data available in a structured format? 2) Is the problem large enough to justify the labelling burden? The first question is whether there are enough suitable and structured data to answer the question. Structured data are much easier to handle and analyse than unstructured data and should always be prioritised. If structured data are found, then their suitability should be tested, for example for completeness and accuracy. In the example, some structured data were available. For instance, crimes are classified as burglaries. These structured data enable the search space to be reduced but do not enable a detailed scan of the variation within burglaries. It is known that information about the variation of interest (outbuilding or not) is not accessible from the structured data. Therefore, PTMs are useful for extracting information from the unstructured data source and for presenting it in a structured manner. Accordingly, in this thesis, burglaries were classified depending on whether they only targetted an outbuilding.

The second consideration has to do with the volume of potential incidents that need to be scanned. All of the studies in this thesis show that PTMs, being a form of supervised learning, require labelled data for fine-tuning. For PTMs to be accurate when applied to the data that were used in this thesis, it was necessary to read and label between 700 and 900 MO texts. This impacted the utility of the PTMs. For the labelling exercise to be efficient, the pool of potential incidents must be sufficiently large. If the area of interest had only had 100 burglaries in the previous year, then the PTMs would not have been efficient. However, if an area, and perhaps a comparison area, have had thousands of burglaries, then it becomes more likely that the PTMs would be efficient.

PTMs are likely to be useful at the scanning stage. Their usefulness depends on there being a gap in the knowledge that is generated from the structured data and sufficiently serious potential problems that would justify the effort of using PTMs (primarily labelling costs). This is predicated on a known problem. There can be occasions on which the exact nature of a problem is not known, as alluded at the start of the section. In that case, PTMs and NLP techniques can be used, but not in the way (supervised) in which they were employed in the studies here. For unknown problems PTMs must be used unsupervised. In the unsupervised case, the machine learning algorithm clusters the data according to the variation that the PTM finds. A similar method was used by Birks et al. (2020). They clustered burglaries without using prior information about the desired themes of the clusters. This highlights a key limitation of the use of PTMs in the way that is explored in this thesis (supervised) – one must know what problem variation one is looking for in order to explore it.

In summary, PTMs can be useful for the scanning phase of the POP process because they allow additional information to be unearthed from unstructured data sources (Study 1a demonstrated how PTMs outperformed current keyword searches). That information can then be used to solve group problems. Once grouped, the problems need to be analysed in order to determine how they occur. This issue forms the subject matter of the next section.

15.2.2 Analyse in Depth

This part of the framework entails arriving at a more complete understanding of the causes of problems. What underlying mechanisms generate a given problem? Although there are some variations between problems, it is important to identify key areas of overlap. POP practitioners must delve deeper into problems than in the scanning phase. They must gain more information about the problems in order to understand developments. This deeper analysis is likely to involve more unstructured data, and PTMs can facilitate its systematic analysis. The set of relevant data sources is likely to be expanded. Although only high-level overviews of the problem may be utilised, the analysis phase is likely to involve work with more detailed and therefore lengthier documents, such as witness statements and other police reports.

This thesis showed that as documents become longer, the ability of PTMs to analyse them efficiently becomes more limited. The structure of PTMs does not allow computations to be scaled linearly with the length of the document. Consequently, the PTM analysis of longer documents requires more computational resources. Study 2 introduced a PTM, Longformer, which is designed for longer texts. The texts in Study 2, although larger than MO texts, were not particularly long, in terms of word count, especially if compared to witness statements. The median length of the police incident logs was 166 words. Witness statements can run to several pages. With each page containing up to 500 words, they are likely to be longer than police incident logs and therefore to require more computational resources. The use of existing PTMs for texts of this length has not been studied extensively. However, one paper from the medical literature (Gao et al., 2021) indicates that current models loose some of their effectiveness when applied to long texts (circa 2,000 words). This loss of effectiveness, coupled with the high computational costs, may mean that, at this stage of their development, PTMs are not suitable for the more detailed work that the analysis phase requires. Other NLP models may be appropriate, depending on the exact nature of the problem and the texts, but that issue is not investigated here. For instance question-and-answering models may allow a more detailed extraction of information.

In short, the analysis phase of the SARA framework is not likely to be the most appropriate for the exploitation of PTMs due to model limitations that have to do with the lengths of texts. Long texts are required in this phase of SARA because of the additional detail required for each incident.

15.2.3 Respond

The response stage is about designing and implementing a response. This phase is about considering the evidence that has been amassed in the course of the two preceding stages and about designing a strategy for eliminating the conditions that cause problems to occur. Ideally, a response should not lean on enforcement activity and should account for previous solutions to similar problems. Unlike the first two stages, the third one does not entail reading similar descriptions of problems and trying to extract information from the texts. Therefore, this phase is unlikely to benefit from the use of PTMs. PTMs are most useful when used to complete repetitive tasks. Once a response has been implemented it should then be assessed to see if it has made the desired impact.

15.2.4 Assessment

The final stage of the POP framework is assessment. The assessment of a POP response determines 1) whether it solved the problem at hand and 2) what mechanism caused it to be effective. Assessment normally revolves around count data and statistical tools that can identify changes. This can be somewhat limiting because the count methodology is constrained by the predetermined categories that the police use to record crime. Relying on count data in this manner can cause variation in crimes within the same classification to be obscured.

Intra-crime variation might mask the success of a POP implementation. For example, a popular response to burglaries is to make targets (typically houses) more difficult to breach. Offenders may then begin to only break into the (less protected) outbuildings or to rely on open windows and doors (i.e., not forcing entry). Neither of these changes in variation would manifest in a typical count-led evaluation strategy because both still constitute burglary. Undoubtedly, however, if offenders change their techniques, then the POP response may be said to have affected them. PTMs can identify this intra-crime variation and thus supplement the count-led assessments of a POP response. It may be difficult to determine what variation one must seek. Therefore, additional consideration of the possible contexts, mechanisms, and observations is required (Pawson & Tilley, 1997) to prepare the PTMs.

PTMs can enrich POP assessments considerably. A PTM can enable a more thorough assessment of intracrime variation with the same set of resources.

15.2.5 PTMs in the SARA framework

The analysis above, which is based on the results from the studies that were presented in this thesis, implies that PTMs can be useful for POP practitioners. The utility of PTMs is most apparent in the initial and the final stages of the framework. In both instances, the PTMs are useful for exploring intra-crime (or intra-problem) variation. At the beginning of the application of the framework, PTMs are useful for categorising similar problems. At the end of the POP cycle, PTMs can be used to explore how criminal activity has changed even if the number of crimes that are classified in the same way remains unchanged. The utility of PTMs in the response phase is not immediately obvious. PTM usage in the assessment phase could be expanded if PTMs are improved so as to work with longer text documents.

The next section reviews the two additional questions that have a bearing on the utility of PTM usage. How is POP implemented by the police forces that hope to use PTMs? Can the labelling burden(i.e. the burden associated with manually labelling text data for the purposes of a supervised model) be reduced through the sharing of PTMs across police forces?

15.3 Additional Considerations

15.3.1 POP implementation

Chapter 3 introduced two broad approaches to the implementations of POP – the generalist and the specialist approach. Under a generalist approach, individual officers are allowed to complete POP cycles. The specialist approach involves the building of specialist capacity within a police force and within the unit that conducts large-scale POP interventions. Can PTMs be used for each type of implementation? Can PTMs support both the generalist and the specialist POP approaches?

Two key factors emerge from the research that was presented on these pages. The two are problem size (i.e., the number of problems to be tackled) and technical capacity. Problem size is important because PTMs require a certain amount of sample data to learn. For example, in the PF1 burglary data, the PTMs required between 700 and 900 MOs texts to learn the classification accurately. The model could then label thousands of crimes, thus saving time. However, if the problem is small, for instance because the area of interest is not large, then the number of problem texts may be insufficient for the training and use of PTMs. Under the generalist approach, which has individual officers complete POP cycles, the number of problems may not be large. Consequently, PTMs may not be useful to such officers. Specialist teams, which might be operating on a larger scale, are more likely to face problems of an appropriate size. Therefore, the efficiencies of PTMs are more likely to be realised by specialist teams.

Secondly, and relatedly, is the resources that are required to utilise a PTM. These resources include the effort that must be expended to label the texts, the know-how that is necessary to use PTMs, and computing power. Generalist implementations may be affected by a lack of resources, whereas specifically resourced teams are more likely to be able to call on the necessary competence and hardware. Some of these issues can be overcome through more accessible tooling. Tooling can automate some of the implementation measures, and cloud-based solutions can supply additional computing power for short projects. However, these tools were not investigated in this thesis.

It is likely that, at least in the short term, PTMs will be more useful to specialised POP units that possess sufficient and appropriate data and the resources that are needed to implement them correctly.

15.3.2 Model sharing

POP is traditionally associated with the development of centres of excellence and the dissemination of best practices. This approach could be transplanted to encompass fine-tuned PTMs. The results from Study 1c demonstrate that models that are trained in one police-force area can be useful in another, albeit with inferior performance. Model sharing could reduce the labelling burden that the use of PTMs entails. Even PTMs that exhibit lower performance are useful because they can be fine-tuned further on the data of the new police force and reach the desired performance level. Therefore, the transfer of models across police forces can be used as a shortcut, effectively reducing the labelling burden.

The models that are shared should target a specific problem. The problems that different police forces must solve are likely to overlap. These overlaps mean that the PTMs that are fine-tuned by one police force are likely to be useful to another police force. A certain amount of documentation must accompany the PTMs. That documentation should define the original classification task of the PTM precisely. For example, in the case of the vehicular theft classification task from Study 1a and Study 1c, the following details would need to be included in the documentation:

- The PTM was only fine-tuned on residential burglary data.
- Only data from 2018/19 was used.
- A motor vehicle included cars, vans , motorbikes and quadbikes. But not mobility scooters.
- The vehicle, not just the car keys, had to be stolen (Some forces are also interested in the targeting of car keys).
- The vehicle had to be removed from the property to classify as stolen.

The intricacies that emerge in the course of the labelling process as cases at the boundary of a classification are revealed indicate that there can be subtle variations in the problems on which a PTM is finetuned. If PTMs are to be shared, these subtleties must be captured and described alongside the models. To ensure that the receiving force do not misinterpret the purpose of the model.

15.4 Implementation Issues

This section investigates some cross-cutting issues that may prevent or delay the use of PTMs in police forces. Most of these issues were mentioned in the previous sections, but they are described here as well for completeness. The implementation issues are explored in three subsections, which concern 1) physical barriers to implementation, such as infrastructure; 2) technical barriers to implementation, which are based largely on gaps in knowledge; and 3) ethical barriers to implementation, which have less to do with the possibility of using PTMs and are more intimately connected to the desirability of their application.

15.4.1 Physical

This section is related to the physical barriers to using PTMs. These barriers emerged primarily from work and discussions with PF2. The physical barriers are generally related to the infrastructure that is required to run the PTMs. These physical barriers come in two forms, namely hardware and software.

PTMs, especially the ones that are used with longer texts, require higher-specification hardware than what is generally available to police analysts. This specialist hardware includes additional computer memory (RAM) and computing power. The required change, to meet the bare minimum requirements, is not dramatic – the costs are unlikely to exceed £1,000 per machine. In short, such an upgrade would be easy to implement if desired.

Upgrading software is more difficult because the police use secure systems. Software is subjected to a rigorous process for preventing cyberattacks and data leaks. The work that is presented in this thesis relied heavily on open-source models and applications from the Internet. At present, they cannot be used on police computers (at least in the UK). Although as demonstrated by this work they can be used in a secure working environment.

There are two overarching solutions to these problems – centralisation and localisation. Centralisation would involve creating a central hub of excellence where the PTMs would run. Police forces would send their data, some of which would be labelled, and their queries to the hub. The central hub would then run the PTMs, conduct explainability and bias checks, and return the results to the police forces. Localisation would entail providing individual analysts with more powerful machines and bespoke software, which is yet to be created, enabling them to run the PTMs and analyse the data. Both solutions have numerous advantages and disadvantages, some of which are explored in the technical section that follows.

15.4.2 Technical knowledge

Technical knowledge is related to the technical know-how that is needed to implement a solution. One of the reasons for using PTMs is that they do not require extensive knowledge. Previously, the main hurdle to utilising similar NLP techniques was the implementation of feature extraction. Since the PTMs are pretrained, feature extraction is no longer necessary. The main effort that must be expended is that of labelling the data, which requires subject-matter expertise that police forces already possess. Knowledge of other technical matters, such as hyperparameter tuning and tests for explainability and bias, can be grasped easily by a competent police analyst. Therefore, the use of PTMs in the manner in which they were employed here entails a technical burden, but police analysts are generally capable of shouldering it, especially if provided with specific training.

Implementing PTMs can be simplified further by automating solutions in order to produce the desired results. Software applications can be built so as to abstract the intricacies of the implementation of these solutions. This abstraction requires more initial effort but would enable PTMs to be used more widely with less training. The interpretation of results, especially results on bias and explainability, would still require subject-matter knowledge, but this is a relatively light burden. As with any abstraction, a decrease in flexibility is to be expected – if a PTM performs poorly, over-reliance on automated applications may make it difficult to modify the model and/or the data.

15.4.3 Ethical

There are two main categories of ethical considerations. Firstly, there are the ethical considerations that have to do with bias and explainability. They were covered in the thesis and are captured, among other issues, by the ALGO-CARE framework (Oswald et al., 2018). The second category has to do with the data that are being analysed. The penultimate section of the data chapter discussed limitations and their implications for bias. The second part of the present section focuses on the impact of those limitations, specifically those of data coverage and information completeness.

Model Implications

The ALGO-CARE framework that was introduced in Chapter 6 allows the leadership of the police to decide whether it would be appropriate to deploy an algorithmic tool such as a PTM. The ethical implications that are derived from ALGO-CARE and which the studies explored are bias and explainability. Algorithmic tools are often biased toward certain segments of the data. This bias can manifest in the resource allocations that these tools might influence. Explainability is important because it generates trust. The model should produce outcomes on the basis of correct information and not on the basis of spurious correlations.

These issues were partly addressed in this thesis through the introduction of methods for the conduct of the analysis and through the presentation of results on both bias and explainability. In both respects, the results, which are limited, are promising. In particular, in the cases in which it was possible to analyse the sex and ethnicity of victims, no evidence of bias was found. The explainability results were also promising. However, there were some issues with the automatically generated text (which reflected how the complaint was made to the please i.e. either electronically or by telephone) that would need to be addressed further.

The investigations that are presented in this thesis are limited. The studies only address one type of crime, one type of incident, and a limited set of classification tasks. In short, this thesis does not present a comprehensive investigation of bias in the use of PTMs. Therefore, bias remains a valid ground for ethical concerns about the use of PTMs with police data. These ethical concerns can be mitigated by conducting bias checks on a case-by-case basis and by utilising the results from the PTMs only if it is established that they do not raise ethical issues.

Data Implications

Chapter 3 introduced two limitations of the data that were used in this thesis. These limitations also extend to the use of PTMs for POP. The first issue is police data coverage. The police do not learn about many crimes. As explained in Chapter 3, the resultant gaps are systematic and not random. Nonrandom gaps mean that if the police only combat the crimes that they are aware of, then their resources are not applied equitably. If a biased process becomes more efficient, it is likely to exacerbate inequality further. If PTMs make POP more efficient, then POP efforts may be directed to crimes for which appropriate textual descriptions are available. The resultant pattern would be nonrandom and would likely causes POP implementations to focus on areas with more complete crime records to the detriment of others.

The second implication, which is related, concerns the completeness of the recorded crime data. If it is known that the recording of crimes is influenced by social and economic factors, then it conceivable that the comprehensiveness of the information that the police receive varies. One might also reasonably expect that the relationship between a police officer and a citizen may influence the amount of information that the latter is prepared to share with the former. Other factors, such as the absence of a common language, may also reduce the quality of crime reports. To the best of the author's knowledge, the completeness of MO descriptions has not been researched. Likewise, the influence of victim characteristics on the completeness of police texts has not been explored. Such studies would be important because PTMs, similarly to other NLP techniques, rely exclusively on the textual descriptions that are presented to them. If those descriptions are biased in any way, then so are the results.

These considerations may affect implementation. As highlighted by the ALGO-CARE algorithm, the police are required to ensure that PTMs are used responsibly. These ethical considerations need not prevent the use of PTMs – the variable quality of police data does not obstruct other crime prevention efforts – but they must be examined in order to ensure that biases are not further exacerbated.

15.5 Conclusion

In summary, it has been shown that PTMs can be useful for POP practitioners. In particular, the PTMs can be used in the scanning phase to search for similar problems and in the assessment phase to understand how intra-crime variation may have changed as a result of a POP response. In each case, the PTMs are used to extract structured information from unstructured text, thus making intra-crime variation easier to quantify. In the near term, PTMs are more likely to be used by specialist POP teams because they tend to face more widespread problems and to possess the resources that are required to efficiently leverage PTMs.

There are physical, technical, and ethical barriers to the use of PTMs. The physical and technical problems can largely be overcome through the provision of additional computational resources and training. The ethical considerations are likely to prove less tractable. More research is needed to ensure that the models do not perpetuate known biases.

The next chapter concludes the thesis by indicating how PTMs can be used more broadly with police data and what other areas of research would benefit from the implementation of PTMs for POP practitioners.

Chapter 16

Conclusions

16.1 Introduction

This is the final chapter of the thesis. It consists of two sections. The first section explores potential further study to the research that was presented in this thesis. There are three main avenues for further study. The first avenue is general next steps, improving the models that were used in this thesis in the way that they were used in this thesis. The second avenue explore a how PTMs can be applied in other ways rather than just classifying passages of texts as done here. The final avenue explores how other text data types might be available to be analysed with PTMs. The second section of this chapter summarises the thesis and presents some concluding thoughts.

16.2 Future Research

Research on the use of NLP models with police data is likely to expand considerably. The field of NLP is growing continuously, and NLP models are continually improving. The application of NLP to police data is and will continue to be dynamic research area as NLP techniques improve. This section is split into three sub-sections. Firstly, there is a section on models. This section focuses broadly on potential improvements to the study that are beyond its present scope. The applications section is a brief description of the manner in which NLP models can be used more widely and the third explores additional data types.

16.2.1 Models

This section focuses on the manner in which future research can improve the models that were presented here. The focus is on the use of PTMs to classify short texts.

Further replication

The studies that were presented here are narrow in scope. Only one type of crime (plus ASB) and three types of classification were examined. Although this was partially replicated across two different police forces, the study should be expanded to include additional crimes, different classifications, and other police forces. The replication of the use cases at different police forces would produce a much more refined understanding of their potential for reuse. If models can be reused across police forces, the labelling burden would decrease – only a single model, rather than 43 separate ones (one for each force in the UK), would need to be produced.

Type

This thesis is based on the BERT model. Other PTMs have been produced since BERT was first released, and they are available for free use. Each of these PTMs has distinct characteristics, capabilities, and focal linguistic areas. By experimenting with different kinds of PTMs, one can discover which one works best for a specific uses case. For example, ROBERTA (Liu et al., 2019), a popular PTM, uses a different method to define the words that are used. It handles previously unseen words more robustly. Police data that contain numerous acronyms or obscure words may be represented more accurately by a model of this kind. Consequently, the classifications may become more accurate. Other models have larger architectures, that is, they enable more parameters to be tuned. Such models can represent more intricate nuances in texts and therefore yield more accurate classifications. Model types are likely to evolve. The identification of the models that are most suitable for police data is likely to be a continuous process.

Hyperparameter tuning

Hyperparameters were introduced in an earlier part of the exposition. Hyperparameters are variables in model formulations that alter the training of the model slightly. An example of a hyperparameter is the number of epochs, that is, the number of instances on which the whole training set is used to train the model. Three epochs were used in this study – the model saw each piece of training data on three separate occasions. The tuning of hyperparameters involves adjusting their values in order to optimise performance. This tuning can be a time-consuming process, but it is important because an appropriate combination of hyperparameters can improve model performance. Hyperparameters were not tuned in the studies that were presented here because the thesis is driven by a desire to use simplified processes for text classification that can be implemented easily by a police force. The results from the thesis indicate that default, that is, untuned, hyperparameters produce satisfactory models. Hyperparameter tuning could lead to more accurate classifications or to lower requirements for labelled data. In any event, hyperparameter tuning produce an improvement in model performance and is thus a suitable avenue for further research.

Outcome weighting

In this research, misclassifications and correct classifications were weighted equally. Therefore, the models were trained to reduce the number of incorrect classifications. However, as explored in the conclusion to Study 2, misclassification is not equal in all instances. For instance, missing a burglary in which a car is stolen may be less desirable than the misclassification of a burglary in which no car is stolen. More vividly, missing a vulnerable victim may be more costly than misclassifying a nonvulnerable victim.

A technique that is called “outcome weighting” is used in machine learning to adjust the importance of different outcomes in classification problems. For example, missing a vulnerable victim might be deemed to be twice as costly as classifying a nonvulnerable victim as vulnerable. This cost function must be built with the end user so that their understanding of the problem and the costs of misclassification can be coded into the training of the model. Typically, this weighting can be encoded into the loss function, changing the training of the model. Alternatively, the model outputs can be used in a more sophisticated way so as to deliver the desired outcome.

Vocabulary

BERT recognises a set of words. This set is called a vocabulary. The benefit of a word being in the vocabulary is that it has a clearly defined numerical representation. If a word is not in the vocabulary, then it is broken down into word pieces until it is recognised. In extreme cases, some words can be classified as “unknown”. Breaking a word into pieces can destroy some of its meaning because it is not represented as a single entity. In texts with many out-of-vocabulary words, the meaning of those words may not be represented accurately. Consequently, the classification models may become less accurate.

There are two ways to overcome this problem. Firstly, the vocabulary of BERT can be extended. The most popular unknown words can be added to the vocabulary, thus preventing words form being broken down into pieces. Secondly, unknown words can be changed to words that are already within the BERT vocabulary. For instance, “untidy” was not recognised by BERT and can be replaced with “messy” or “not tidy”, which are both recognised by BERT.

Overall, aligning texts and BERT vocabularies can help to improve the performance of PTMs on specific tasks by increasing the extent to which the models understand the domains in question.

Pre-train

As mentioned previously, there are two stages to utilising a BERT model. There is the pretraining element, which is resource intensive and equips the model with a general understanding of language, and there is also the pretraining that is conducted for each specific task. Pretraining was not completed in the studies that are presented here, but scholars in other domains who have accessed sufficient data have completed it. When this pretraining is conducted, it is possible to build new variants of BERT. For instance, Legal-BERT was built to understand legal documents (Chalkidis, Fergadiotis, Malakasiotis, Aletras, & Androutsopoulos, 2020). Another variation, which is trained on medical data, is called med-BERT (Rasmy, Xiang, Xie, Tao, & Zhi, 2020).

Pretraining a BERT model on police data, perhaps exclusively MO data from several different forces, in order to produce an MO-BERT would be an interesting avenue for future research. Given the successes that have been achieved in other domains, this new model is likely to perform better at classifying MO texts than the regular BERT. This approach may save time and resources that would otherwise need to be allocated to fine-tuning for each additional task as it would possess a superior understanding of the domain-specific language from the outset.

This section demonstrated that there are a number of interesting avenues for additional research that would enhance the classification work that was described in this thesis. The next section takes this further by exploring other NLP techniques that are not intended for the classification of text passages and their potential for enhancing POP.

16.2.2 Applications

This next section introduces additional applications of PTMs above the classification type that was used in this research. Additional applications are important because they allow information to be extratcted from the data in different ways or enable access to the key elements of the data (e.g. summarisation).

Question and Answer

Question answering (Q&A) is an NLP task that involves using a PTM to answer questions that are posed in natural language, given a text which contains the answer. For example, a police officer may obtain one or more documents. These documents are then submitted to the PTM along with a question. The PTM generates a response by only using the text of the documents. Q&A systems can be designed to answer a wide range of questions, including factual questions, questions about definitions, and queries about individuals. However, they have not been trialled with crime documents, and their performance in this domain is unknown. This may be more useful when the incidence of a crime is low or when a specific response, rather than a binary classification, is needed.

Named Entity Recognition

Named entity recognition (NER) is another NLP task that is based on classification. Instead of classifying a passage of text, it classifies every word within it. The typical task is to extract the names of organisations, people, and places from a passage. The PTM labels each word as either “nothing”, “a person”, “an organisation“, or “a place”. For example, in the sentence ”Boris Johnson was born in New York on 19 June 1964”, the named entities are “Boris Johnson”, “New York”, and “19 June 1964”. When words are identified positively, they can be extracted from the text. The PTM uses both the word and the context in which it appears to arrive at a classification. Therefore, names or places that do not feature in the training material can still be extracted correctly because they are used in similar contexts. In policing, the words of interest may not be people or places, but, for example, weapons that are used in assaults.

Summarisation

Text summarisation involves the production of short synopses from longer documents or collections of documents. The idea is to retain the most important information from the original documents so that the summarisation can be read in isolation. This task is well understood in the NLP field, but it is hard to generalise across different language domains because 1) it is inherently difficult to measure an appropriate summary and because 2) different facts matter in different domains. Since the quality of a summary is difficult to quantify, human intervention in the modelling process is necessary, which makes text summarisation more resource intensive than other NLP tasks. The importance of text summarising for the police domain is that it could enable cases to be reviewed more rapidly and easily.

This section introduced three methods beyond classification that could be leveraged to support the analysis of free text data. The next section moves beyond models and explores that data types that can be analysed.

16.2.3 Data

Document types

Police forces use many text documents that are not MO descriptions and incident logs. Police forces generate a large volume of text data when they conduct their operations. These data include case summaries, witness reports, communication logs, and arrest reports. The underlying characteristics of these documents vary and, potentially, so do the techniques that are applicable to them. As mentioned previously, longer documents are difficult to analyse with PTMs because of the additional memory that is required to track the entire document. Shorter documents, such as communication logs, may contain large numbers of abbreviations or irregular grammar, especially if they have been recorded hastily or if they are verbatim reports. These different document types therefore entail different challenges that must be overcome in different ways. Researching PTMs with the different document types would be conducive to a more extensive understanding of the effectiveness of PTMs in the police domain.

Languages

The studies that were presented here concern texts in English, but similar models exist for other languages e.g. (Scao et al., 2022). Translation models make it possible to translate non-English text into English in order to use English-language models. Further research on the use of non-English models to prove that similar tasks are conceivable in languages other than English would also be useful in proving that the approaches that were explored here can be used widely.

Bias

The results from the bias investigations that were presented here are encouraging, but they only reflect one part of the journey from data creation to model output. In particular, the studies that were presented here focus on algorithmic bias. To the best of the author's knowledge, little is known about the biases that affect the comprehensiveness of textual crime records. This section of the data journey merits additional research.

16.2.4 Explainability

Although explainability was introduced and explored in the thesis, the visualisations that were provided were not tested formally for effectiveness. Explainability, as outlined in Chapter 6, is specific to audiences. Accordingly, visualisations and methods should be tested with all intended audiences. This testing should include members of the public (e.g., victims who have suffered crimes that may be classified), the police officers who use the outputs, and the individuals who authorise the use of the models.

16.2.5 Summary

In summary, there are many potential avenues for research, particularly those that have been shown to work in other domains. However, two general problems may restrict the use of NLP with police data. Firstly, long texts are problematic because computational requirements increase quadratically with text length – very long texts require considerably more computational power and memory than what is generally feasible. Secondly, around 1,000 labelled texts were required for each problem in this thesis. In some instances, resource demands may be disproportionate to the gravity of the problem that has to be solved. These challenges, however, are being tackled actively, and it is likely that the restrictions that were outlined will become less stringent in the future. Consequently, NLP techniques will become applicable to a wide variety of police data and police requirements for interrogating free-text data.

16.3 Concluding Remarks

This thesis set out to determine whether modern NLP methods, namely PTMs, can be employed effectively within the POP methodology. The author hoped that the analytical burden of the POP process would be reduced. The analytical requirements were previously thought to obstruct the successful completion of POP projects. The use of PTM was explored by exploring intra-incident variation experimentally in relation to descriptions of burglary and ASB. In each case, intra-incident variation was specified by using PTMs to arrive at classifications on the basis of a predetermined characteristic. The robustness of the results was explored by reference to model transparency and bias. The results for burglary were also replicated with data from different police forces.

The results of the experiments were then analysed in the context of the POP cycle, and they were found to be particularly useful in the initial (scanning) and the latter (Assessment) phases of the POP cycle. PTMs can reduce the analytical burdens of that cycle. Issues of implementation were also discussed, and areas for future work were explored. In conclusion, police forces can benefit considerably from the use of NLP techniques and models. PTMs, in particular, are highly useful for extracting information from free-text material. This information can then be used for a variety of purposes, including the formulation of crime prevention strategies.

As technologies advance there is considerable scope for the ways in which PTMs and other aligned technologies may deliver new efficiencies for policing. As this thesis finishes there have been huge advances in NLP technologies typified by the release of Chat-GPT (Vallance, 2022). Although not perfect these new technologies are likely to allow police forces to access their data in a much more efficient manner. Making police force themselves much more efficient. In conclusion there is a significant possibility that NLP can have a real impact on a police forces ability to combat crime.

Bibliography

Adderley, R., & Musgrove, P. (2003). Modus operandi modelling of group offending: A data-mining case study. *International Journal of police science & management*, *5*(4), 265–276.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International journal of social research methodology*, *8*(1), 19–32.

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115.

Avison, D. E., Lau, F., Myers, M. D., & Nielsen, P. A. (1999). Action research. *Communications of the ACM*, *42*(1), 94–97.

Babuta, A., Oswald, M., & Rinik, C. (2018). *Machine learning algorithms and police decision-making* (tech. rep. No. Whitehall report 3-18). Royal United Services Institute. London.

Bache, R., Crestani, F., Canter, D., & Youngs, D. (2010). A language modelling approach to linking criminal styles with offender characteristics. *Data & Knowledge Engineering*, *69*(3), 303–315.

Bachenko, J., Fitzpatrick, E., & Schonwetter, M. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. (Vol. 1, pp. 41–48). cited By 42. doi:10.3115/1599081.1599087

Basilio, M., Brum, G., & Pereira, V. (2020). A model of policing strategy choice: The integration of the latent dirichlet allocation (lda) method with electre i. *Journal of Modelling in Management*, *15*(3), 849–891. cited By 1. doi:10.1108/JM2-10-2018-0166

Basilio, M., Pereira, V., & Brum, G. (2019). Identification of operational demand in law enforcement agencies: An application based on a probabilistic model of topics. *Data Technologies and Applications*, *53*(3), 333–372. cited By 2. doi:10.1108/DTA-12-2018-0109

Baumer, E. P. (2002). Neighborhood disadvantage and police notification by victims of violence. *Criminology*, *40*(3), 579–616.

Beltagy, I., Peters, M. E., & Cohan, A. (2020a). Longformer: The long-document transformer. doi:10.48550/ARXIV.2004.05150

Beltagy, I., Peters, M. E., & Cohan, A. (2020b). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150.*

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Birks, D., Coleman, A., & Jackson, D. (2020). Unsupervised identification of crime problems from police free-text data. *Crime Science*, *9*(1), 18. doi:10.1186/s40163-020-00127-4

Boulton, L., McManus, M., Metcalfe, L., Brian, D., & Dawson, I. (2017). Calls for police service: Understanding the demand profile and the uk police response. *The Police Journal*, *90*(1), 70–85. doi:10.1177/0032258X16671032. eprint: https://doi.org/10.1177/0032258X16671032

Bowers, K. J., & Johnson, S. D. (2004). Who commits near repeats? a test of the boost explanation. *Western Criminology Review*, *5*(3).

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice quarterly*, *31*(4), 633–663.

Brown, J., & Sturge, G. (2021). *Tackling anti-social behaviour.*

Buil-Gil, D., Moretti, A., & Langton, S. H. (2021). The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology*, 1–27.

Campedelli, G. M. (2019). Where are we? using scopus to map the literature at the intersection between artificial intelligence and crime. *Using Scopus to Map the Literature at the Intersection Between Artificial Intelligence and Crime (December 23, 2019).*

Castelli, V., & Cover, T. M. (1995). On the exponential value of labeled samples. *Pattern Recognition Letters*, *16*(1), 105–111.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. doi:10.48550/ARXIV.2010.02559

Chen, C.-h., & Chi, Y.-P. J. (2010). Use text mining approach to generate the draft of indictment for prosecutor. In *Pacis* (p. 23).

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1–13.

Chohlas-Wood, A., & Levine, E. (2019). A recommendation engine to aid in identifying crime patterns. *INFORMS Journal on Applied Analytics*, *49*(2), 154–166.

Chollet, F., & Allaire, J. (2018). *Deep learning with r*. Manning.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163.

Clarke, R. V. (1995). Situational crime prevention. *Crime and justice*, *19*, 91–150.

Clarke, R. V. (2016). Situational crime prevention. In R. Wortley & M. Townsley (Eds.), *Environmental criminology and crime analysis* (2nd ed., Chap. 13, pp. 286–303). Taylor & Francis.

Clarke, R. V. G. (1997). *Situational crime prevention*. Criminal Justice Press Monsey, NY.

Clarke, R., & Eck, J. (2003). Becoming a problem-solving crime analyst in 55 steps. *London: Jill Dando*.

Cocx, T. K., & Kosters, W. A. (2006). A distance measure for determining similarity between criminal investigations. In *Industrial conference on data mining* (pp. 511–525). Springer.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American sociological review*, 588–608.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eck, J. (2003). Police problems: The complexity of problem theory, research and evaluation. *Crime prevention studies*, *15*, 79–114.

Eck, J. E., & Spelman, W. (1987). Problem-solving: Problem-oriented policing in newport news.

Eisenstein, J. (2018). *Natural language processing*. MIT Press.

Fairness and effectiveness in policing: The evidence. (2004). "National Academy of Sciences". doi:10.17226/10419

Farrell, G., Laycock, G., & Tilley, N. (2015). Debuts and legacies: The crime drop and the role of adolescence-limited and persistent offending. *Crime Science*, *4*(1), 1–10.

Felson, M. (2016). The routine activity approach. In R. Wortley & M. Townsley (Eds.), *Environmental criminology and crime analysis* (2nd ed., Chap. 4, pp. 87–97). Taylor & Francis.

Felson, M., & Clarke, R. V. (1998). Opportunity makes the thief. *Police research series, paper*, *98*, 1–36.

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)* (pp. 363–370).

Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., ... Tourassi, G. (2021). Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3596–3607. doi:10.1109/JBHI.2021.3062322

Goldfarb-Tarrant, S., Marchant, R., Sánchez, R. M., Pandya, M., & Lopez, A. (2020). Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Goldstein, H. (1990). *Problem-oriented policing*. Temple University Press. Retrieved from https://books.google.co.uk/books?id=0hPEQgAACAAJ

Goldstein, H. (1979). Improving policing: A problem-oriented approach. *Crime and Delinquency*, *25*(2), 236–258. Retrieved from http://search.proquest.com/docview/1308274554/

Goldstein, H. (2018). On problem-oriented policing: The stockholm lecture. *Crime Science*, *7*(1), 1, 9.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. *Computer Science Review*, *29*, 21–43.

Gray, C. S. (1999). *Modern strategy*. Oxford: Oxford University Press.

Guerette, R. T., & Bowers, K. J. (2009). Assessing the extent of crime displacement and diffusion of benefits: A review of situational crime prevention evaluations. *Criminology*, *47*(4), 1331–1368.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

Haleem, M., Han, L., Harding, P., & Ellison, M. (2019). An automated text mining approach for classifying mental-ill health incidents from police incident logs for data-driven intelligence. (Vol. 2019-October, pp. 2279–2284). cited By 0. doi:10.1109/SMC.2019.8914240

Halford, E., Dixon, A., & Farrell, G. (2022). Anti-social behaviour in the coronavirus pandemic. *Crime Science*, *11*(1). doi:10.1186/s40163-022-00168-x

Halford, E., Dixon, A., Farrell, G., Malleson, N., & Tilley, N. (2020). Crime and coronavirus: Social distancing, lockdown, and the mobility elasticity of crime. *Crime science*, *9*(1), 1–12.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(3), 139–154.

Helbich, M., Hagenauer, J., Leitner, M., & Edwards, R. (2013). Exploration of unstructured narrative crime reports: An unsupervised neural network and point pattern analysis approach. *Cartography and Geographic Information Science*, *40*(4), 326–336. cited By 26. doi:10.1080/15230406.2013.779780

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*(4), 811–866.

Hinkle, J. C., Weisburd, D., Telep, C. W., & Petersen, K. (2020). Problem-oriented policing for reducing crime and disorder: An updated systematic review and meta-analysis. *Campbell Systematic Reviews*, *16*(2), e1089.

Home Office. (2020). Crime recording general rules. [Online; accessed 20-September-2020]. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/913721/count-general-sep-2020.pdf

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, *2*(4), 100241.

Hwang, Y., Zheng, L., Karystianis, G., Gibbs, V., Sharp, K., & Butler, T. (2020). Domestic violence events involving autism: A text mining study of police records in new south wales, 2005-2016. *Research in Autism Spectrum Disorders*, *78*. cited By 0. doi:10.1016/j.rasd.2020.101634

Jeff Larson, J. A. (2016). How we analyzed the compas recidivism algorithm. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930–1938.

Johnson, S. (2016). Crime mapping and spatial analysis. In R. Wortley & M. Townsley (Eds.), *Environmental criminology and crime analysis* (2nd ed., Chap. 10, pp. 199–223). Taylor & Francis.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Karystianis, G., Adily, A., Schofield, P. W., Greenberg, D., Jorm, L., Nenadic, G., & Butler, T. (2019). Automated analysis of domestic violence police reports to explore abuse types and victim injuries: Text mining study. *Journal of medical Internet research*, *21*(3), e13067.

Karystianis, G., Adily, A., Schofield, P., Knight, L., Galdon, C., Greenberg, D., ... Butler, T. (2018). Automatic extraction of mental health disorders from domestic violence police narratives: Text mining study. *Journal of medical internet research*, *20*(9), e11548.

Keyvanpour, M., Javideh, M., & Ebrahimi, M. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. (Vol. 3, pp. 872–880). cited By 39. doi:10.1016/j.procs.2010.12.143

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807.*

Krause, S., & Busch, F. (2019). New insights into road accident analysis through the use of text mining methods. cited By 0. doi:10.1109/MTITS.2019.8883343

Krishnamurthy, R., & Kumar, J. S. (2012). Survey of data mining techniques on crime data analysis. *International Journal of Data Mining Techniques and Applications*, *1*(1), 47–49.

Ku, C.-H., & Leroy, G. (2013). Automated crime report analysis and classification for e-government and decision support. (pp. 18–27). cited By 0. doi:10.1145/2479724.2479732

Kuang, D., Brantingham, P. J., & Bertozzi, A. L. (2017). Crime topic modeling. *Crime Science*, *6*(1), 12.

Kumar, E. (2011). *Natural language processing.* IK International Pvt Ltd.

Langton, S., Dixon, A., & Farrell, G. (2021). Six months in: Pandemic crime trends in england and wales. *Crime science*, *10*(1), 1–16.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Maguire, E. R., Uchida, C. D., & Hassell, K. D. (2015). Problem-oriented policing in colorado springs: A content analysis of 753 cases. *Crime & Delinquency*, *61*(1), 71–95.

Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, *41*(4), 701–707.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge university press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).

Marshall, B., & Townsley, M. (2006). *Needles or needless?* Jill Dando Institute, UCL. Citeseer.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC medical research methodology*, *10*(1), 1–16.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, *3*(4), 235–244.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Narayanan, A., & Felten, E. W. (2014). No silver bullet: De-identification still doesn't work.

Nayak, A., Timmapathini, H., Ponnalagu, K., & Gopalan Venkoparao, V. (2020). Domain adaptation challenges of BERT in tokenization and sub-word representations of out-of-vocabulary words. In *Proceedings of the* *first workshop on insights from negative results in nlp* (pp. 1–5). doi:10.18653/v1/2020.insights-1.1

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the durham hart model and 'experimental'proportionality. *Information & Communications Technology Law*, *27*(2), 223–250.

Paiva, P. Y. A., Moreno, C. C., Smith-Miles, K., Valeriano, M. G., & Lorena, A. C. (2022). Relating instance hardness to classification performance in a dataset: A visual approach. *Machine Learning*, *111*(8), 3085–3123.

Pandey, R., & Mohler, G. (2018). Evaluation of crime topic models: Topic coherence vs spatial crime concentration. (pp. 76–78). cited By 1. doi:10.1109/ISI.2018.8587384

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* sage.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2009). A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5633 LNAI*, 247–260. cited By 17. doi:10.1007/978-3-642-03067-3_20

Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., & Dedene, G. (2009). Gaining insight in domestic violence with emergent self organizing maps. *Expert Systems with Applications*, *36*(9), 11864–11874. cited By 9. doi:10.1016/j.eswa.2009.04.027

Poelmans, J., Van Hulle, M., Viaene, S., Elzinga, P., & Dedene, G. (2011). Text mining with emergent self organizing maps and multi-dimensional scaling: A comparative study on domestic violence. *Applied Soft Computing Journal*, *11*(4), 3870–3876. cited By 10. doi:10.1016/j.asoc.2011.02.026

Prokofyev, R., Demartini, G., & Cudré-Mauroux, P. (2014). Effective named entity recognition for idiosyncratic web collections. In *Proceedings of the 23rd international conference on world wide web* (pp. 397–408).

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.".

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2020). Med-bert: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. doi:10.48550/ARXIV.2005.12833

Ratcliffe, J. H., & McCullagh, M. J. (1998). Aoristic crime analysis. *International Journal of Geographical Information Science*, *12*(7), 751–764.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the vldb endowment. international conference on very large data bases* (Vol. 11, *3*, p. 269). NIH Public Access.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866.

Rogerson, M. (2016). *The utility of applying textual analysis to descriptions of offender modus operandi for the prevention of high volume crime* (Doctoral dissertation, University of Huddersfield).

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., . . . Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100.*

Schraagen, M., & Bex, F. (2019). Extraction of semantic relations in noisy user-generated law enforcement data. (pp. 79–86). cited By 1. doi:10.1109/ICOSC.2019.8665497

Scott, M. S., & Clarke, R. V. (2020). *Problem-oriented policing: Successful case studies.* Routledge.

Scott, M., Eck, J., Johannes, K., & Goldstein, H. (2016). Problem-oriented policing. In R. Wortley & M. Townsley (Eds.), *Environmental criminology and crime analysis* (2nd ed., Chap. 11, pp. 227–258). Taylor & Francis.

Scott, M., & Kirby, S. (2012). Implementing pop: Leading, structuring and managing a problem-oriented police agency. Community Oriented Policing Services, US Department of Justice.

Seo, S., Chan, H., Brantingham, P. J., Leap, J., Vayanos, P., Tambe, M., & Liu, Y. (2018). Partially generative neural networks for gang crime classification with partial information. In *Proceedings of the 2018 aaai/acm conference on ai, ethics, and society* (pp. 257–263).

Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Sheard, E. J. (2020). *Developing a combined risk model for the prediction of temporally clustered offences* (Doctoral dissertation, University of Leeds).

Sidebottom, A., Bullock, K., Armitage, R., Ashby, M., Clemmow, C., Kirby, S., . . . Tilley, N. (2020). Problem-oriented policing in england and wales 2019.

Sidebottom, A., Kirby, S., Tilley, N., Armitage, R., Ashby, M., Bullock, K., & Laycock, G. (2020). Implementing and sustaining problem-oriented policing: A guide.

Spiegelhalter, D. (2019). *The art of statistics: Learning from data*. Penguin UK.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 197–207). doi:10.18653/v1/K18-2020

Tarling, R., & Morris, K. (2010a). Reporting Crime to the Police. *The British Journal of Criminology*, *50*(3), 474–490. doi:10.1093/bjc/azq011. eprint: https://academic.oup.com/bjc/article-pdf/50/3/474/1281745/azq011.pdf

Tarling, R., & Morris, K. (2010b). Reporting crime to the police. *The British Journal of Criminology*, *50*(3), 474–490.

Tilley, N. (2010). Whither problem-oriented policing. *Criminology and Public Policy*, *9*(1), 183, 195.

Vallance, C. (2022). Chatgpt: New ai chatbot has everyone talking to it. BBC. Retrieved from https://www.bbc.co.uk/news/technology-63861322

van de Putte, X., Oling, P., & Schakel, J.-K. (2009). In search for intelligence: Automatically estimating the implicitness of police officers' observation messages, an ongoing action research. (pp. 425–432). cited By 0. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84873880119&partnerID=40&md5=2e331d73051e53ab4f21994425ba9411

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1–7). IEEE.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461.*

Weisburd, D., Telep, C. W., Hinkle, J. C., & Eck, J. E. (2010). Is problem-oriented policing effective in reducing crime and disorder? *Criminology and Public Policy*, *9*(1), 139, 172.

Weisel, D. L. (2016). Analyzing repeat victimization.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 9.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th). Morgan Kaufmann.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771.*

Wortley, R. (2010). Critiques of situational crime prevention. In B. S. Fisher et al. (Eds.), *Encyclopedia of victimology and crime prevention* (Vol. 1). Sage Publications, Inc.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., . . . Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 1–21). doi:10.18653/v1/K18-2001