USING ARTIFICIAL INTELLEGENCE TO IDENTIFY PERPETRATORS OF TECHNOLOGY FACILITATED COERCIVE CONTROL.

Home Office Domestic Abuse Perpetrators Research Fund

Tirion Havard, Nonso Nnamokon, Chris Magill, Cyndie Demeocq, Jack Procter, Denise Harvey, and Vanessa Bettinson

havardt@lsbu.ac.uk

Project partners and contact information

LSBU

- Investigator Name
- Address
- Tel
- E-mail

DMU

- Investigator Name
- Address
- Tel
- E-mail

BU

- Investigator Name
- Address
- Tel
- E-mail

EHU

- Investigator Name
- Address
- Tel
- E-mail

Table of Contents

Glossary of Abbreviations

Abbreviation	Name
AI	Artificial Intelligence
CJS	Criminal Justice System
CPS	Crown Prosecution Service
HMICFRS	HM Inspectorate of Constabulary Fire and Rescue Service
IOPC	Independent Office for Police Conduct
LSBU	London South Bank University
ML	Machine learning
NLP	Natural Language Processing
NPCC	National Police Chief's Council
ONS	Office for National Statistic
TFCC	Technology Facilitated Coercive Control
VAWG	Violence Against Women and Girls
WP	Work Package

Acknowledgments

This feasibility study is the result of support and collaboration with individuals from a range of organisations across different sectors all of whom who share a commitment to ending domestic abuse.

First and foremost, a special and extended thank you to goes to all the survivors who took the time to complete the survey and share their views about police using Artificial Intelligence in identifying perpetrators of domestic abuse.

Thank you to all the survivor agencies who promoted the survey across their networks. A special shout out to Rose Ssali CEO of SAWN (Support and Action Women's Network) who promoted participation and engagement of Black and Ethnic Minority Women and offered support sessions to survivors of domestic abuse.

Further gratitude is given to Jo Gough, CEO of RISE (an independent, Brighton-based charity that helps people affected by domestic abuse) for sharing the survey with their survivors and other organisations in their networks. We are grateful too, to Salus, who offer innovative intervention and mentoring to children, young people and families including those who experience violence and abuse, and who put us in touch with Rising Sun, a Kent based domestic violence and abuse service.

Thank you to Viktorija Zdanoviciute from the Health Education Partnership who shared the survey with professionals across health and education sectors. Thanks also to Mark Brooks from Mankind, who ensured the participation of male survivors of domestic abuse.

Thanks to colleagues in the statutory sector, especially Sophie Linden, Deputy Mayor for Policing and Crime at MOPAC, who supported the research from the outset and secured connections within the Metropolitan police. Thanks to each member of staff from the different courts and judiciary who supported us in negotiating the process, responding swiftly to queries and emails and supplying the transcripts. We also appreciate the speedy responses of the transcription agencies.

A special shout out to Matthew Cornish, from Essex Police for his time, support and invaluable advice about the potential of AI in the policing of domestic abuse. Last but not least, thanks to all the officers who took the time to complete the survey and share your thoughts.

The acknowledgement would not be complete without mentioning the Home Office's Perpetrators Programme without whose funding the outcomes of this research project, including this report would not have been possible. The views reflected in this research are not necessarily those of the Home Office

Executive Summary

In 2019, the government publicly acknowledged that the Criminal Justice System is failing victim/survivors of rape and sexual assault resulting in an erosion of public trust and confidence. The abduction, rape, and murder of Sarah Everard by a serving police officer started a series of controversies that served to decrease public trust further. Serious sexual offences are also taking the longest time on record to go through Crown Courts in England and Wales, with the time from the first Crown Court hearing to the end of a case averaging nine months.

The government recognises that the volume of digital data and the length of time it takes to analyse it are significant factors in these delays and that they undermine police investigations and the prosecution process. Police forces report being overwhelmed by the exponential growth in the volume of digital evidence, with over 20,000 digital devices waiting to be processed. Victim/survivors of domestic abuse are also waiting up to four and a half years for the police to return their phones following an investigation. This coupled with victim/survivors feelings of being 'digitally strip searched' (i.e., police interest in seemingly irrelevant information such as search histories relating to shopping or holidays) is contributing to high numbers of survivor/victims withdrawing from cases.

This study is one of the 21 projects funded by the Home Office for research on perpetrators of domestic abuse. It is interested in a specific form of domestic abuse known as Technology Facilitated Coercive Control (TFCC) and focussed on the digital communication between (alleged) perpetrators and victim/survivors held on mobile phones. The purpose of this feasibility study was twofold,

i. to test the viability of an Artificial Intelligence (AI) programme to identify () perpetrators (including alleged perpetrators) of domestic abuse using digital communications held on mobile phones
ii. to examine police and victim/survivor attitudes towards using AI in police investigations.

Using digital conversations extracted from court transcriptions where TFCC was identified as a factor in the offending, the research team tested data sets built on different methods and techniques of AI. Natural Language Processing (NLP) tools, a subfield of AI, were also tested for their speed and accuracy in recognising abusive communication and identifying and risk assessing perpetrators of TFCC.

Conscious of national concern about policing practices relating to Violence Against Women and Girls and that any AI programme would be futile without the co-operation of both the police and the public, two online surveys were devised to measure opinion. The first sought insight into the attitudes of victim/survivors, viewed as experts in domestic abuse, about using AI in police investigations. The second involved the police and questioned their views of using AI in this way.

Organisations who support victim/survivors of domestic abuse and who are known to the research team were approached for their help recruiting victim/survivor participants. To the team's knowledge, this is the first time the views of survivor/victims about the role of AI in police investigations have been sought. Individual police officers or those with connections to the police service were approached inviting them to complete the questionnaire. These organisations and individual participants were also asked to distribute the link to the survey amongst their wider networks. The link was also posted and promoted at regular intervals on Twitter and LinkedIn. As an incentive, victim/survivors of domestic abuse were offered the opportunity to enter a draw for the chance to win a £100 voucher.

A total of 81 victim/survivors from diverse demographics took part in the survey. Results showed that 70% victim/survivors of domestic abuse were willing to share their digital data with the police if AI technology was used. Victim/survivors' feelings of being 'digitally strip searched' was less clear as the responses were more evenly distributed. Comments in the text boxes suggest that victim/survivors of domestic abuse are curious about how AI can be used to help police with their enquiries but have concerns about the bias of such a programme which is, at least in part, linked to a mistrust of the police. Victim/survivors were also aware of the importance of understanding TFCC within a wider context and were unclear as to the programme's ability to do this. More qualitative research is required to gain an in-depth understanding of survivor/victims concerns and hopes for using AI in police investigations in the future.

AI's ability to understand the digital data within the wider context was echoed by some of the 28 police staff who participated in this survey. The issue of bias by this technology was also an issue. Research shows that concern relating to AI bias is often misplaced as it is the data that is subject to bias, not the programme itself. To mitigate against this potential bias, further exploration is required utilising larger data sets. The result of both surveys' also suggests that educating the public to dispel some of the myths around AI technology would be beneficial.

Domestic abuse cases involving TFCC were identified from newspapers and public databases. Six court transcripts were obtained, the digital communication between (alleged) perpetrators and the victim/survivors was removed, anonymised, and entered onto a data base. This provided a usable dataset of 219 messages. Because this research focussed on understanding the behaviour of (alleged) perpetrators of TFCC only communication threads of (alleged) perpetrators were used. This provided a total of 250 relevant messages which were enriched with an additional 242 perpetrator messages obtained from online repositories. Data instances that represent the absence of coercive and controlling behaviour were retrieved from twitter, bringing the total number of messages used in this research to 1012.

Three classifiers (software systems that process text data at scale) were used in this study namely Random Forest, SVM Linear and RBF. All were trained with embeddings from BERT, GPT2, GloVe and Word2Vec. Results showed the technologies are both fast and accurate in predicting perpetrators of domestic abuse. Based on these encouraging findings further research is necessary, with larger data sets, to train models to have an in-depth understanding of TFCC and test its application to diverse real-world scenarios.

This research has tested the feasibility of AI technology to address government concerns about the rate of convictions relating to cases of sexual and domestic abuse. Findings indicate public support (albeit cautious) on behalf of police and victim/survivors for using AI in police investigations. Early results suggest that this technology would quicken the police's ability to process digital data, cut down on the length of time they hold victim/survivor phones, limit delays in court processing and reduce the number of victim/survivors of TFCC who withdraw from cases. Further research is required to test the generalisability of this project and determine how it could best be used to increase the efficiency and effectiveness of the Criminal Justice System when dealing with domestic abuse cases.

.

Introduction

In February 2022, the Home Office awarded £1.5 million to 21 projects designed to build on existing research and address current gaps in knowledge on domestic abuse and identifying perpetrators. This report presents the findings from one of these projects. The project, a feasibility study, was developed by a consortium of academics from different disciplines and institutions (London South Bank University, Edgehill University, the University of Brighton, and De Montfort University). It makes a significant contribution to the use of Artificial Intelligence (AI) to assist decision making within the criminal justice system in domestic abuse cases including the views of survivors/victims and police staff in using this as part of police enquiries. Specifically, our study tests the suitability of natural language processing (NLP) methods as a tool to identify and risk assess (alleged) perpetrators of domestic abuse.

The Domestic Abuse Act 2021, s.1, created a statutory definition of domestic abuse which consists of any incident or pattern of incidents of controlling, coercive or threatening behaviour, violence or abuse between people who are or have been intimate partners or family members. This definition applies to all individuals aged 16 or over regardless of gender or sexuality. The government were keen to emphasise that domestic abuse is not just physical violence, but can also involve emotional, controlling, or coercive behaviours, and economic abuse (Home Office 2022)

Coercive control, a form of domestic abuse (Home Office 2015), indicates a high risk of future physical violence (Felson & Messner, 2014). It consists of a range of behaviours including threats and intimidation, which are designed to harm punish and frighten the victim (Stark, 2007; Woman's Aid 2019). It is likely to impact on every aspect of a victims life (Hamberger *et al*., 2017) but, when taken in isolation can appear insignificant to the outsider (Williamson, 2010). Take for example a remark made by a perpetrator that is viewed by others as gentle banter but within the context of abusive relationships may be a warning to the victim not to cross the line (Stark 2007, Wiener, 2017).

The increasing use of technology in facilitating coercive control has been identified in recent research studies (Havard and Lefevre, 2020). It can involve surveillance, which is an integral part of coercive control, with Stark (2007, p257) describing it as '*almost universal in abusive relationships'*. The role of technology in the monitoring and control of victim/survivors was highlighted in a recent report by Refuge (2020) who found that 72% of women accessing their services said technology had been integrated into their abuse. Examples of this 'Technology Facilitated Coercive Control'(TFCC) include monitoring/controlling through bombarding with texts and phone/video calls, threatening or intimidating via social media, gaining access to women's personal and home devices, online accounts, and children's toys ((Dragiewitz et al., 2018; Woodlock and Harris, 2020). Mobile phones are an established gateway to such abuse, their portability and diverse capabilities are manipulated by perpetrators to develop strategies of 'agile technological surveillance', (Havard and Lefevre, 2020, p224) which allow them to track and monitor their partners in various ways, whilst on the go and irrespective of physical proximity. The current study focuses on the analysis of the digital communication between intimate partners that is contained in textual communication channels such as text messages in mobile phone, email, mobile phone etc. relational aspects of patterns of abuse against intimate partners

AI specifically Machine Learning (ML) makes use of programming languages and algorithms to build predictive or decisional models. ML offers methods for data analysis and computer understanding as the algorithms employed allow the model to learn from historical data which then can be used for prediction on new data (Esposito, D. & Esposito, F., 2020). Natural Language Processing (NLP) is a subfield of ML which focuses on the understanding of text. This analysis is done by using vector representation techniques, where words are converted into digits that are readable by the machine and enable the computer to understand the context of sentences and paragraphs (Pilehvar & Camacho-Collados, 2020).

This project combines ML computational linguistics, and NLP techniques, to develop an enhanced tool that understands natural language for the analysis of sentiments, attitudes, and emotions expressed in the communication between victims and perpetrators of domestic abuse. Text mining and NLP techniques have been used (albeit sparingly) in the broad area of criminal justice (Elyezjy & Elhaless, 2015; Pandey, 2020; Pinho et al., 2017; Riya & Gandotra, 2016; Xu, 2021), including to assist judges and magistrates with bail applications (Kleinberg et, al 2018). However, no existing study has applied this approach to the analysis of TFCC in domestic abuse cases. Early identification of perpetrators and accurate risk assessment tools are essential to ensure effective criminal justice sanctions are taken against perpetrators in domestic abuse cases and to stop them moving from one victim to the next. In 2019 there were 17,616 offences of coercive control recorded by police with only 1,177 (fewer than 7%) resulting in prosecution (ONS, 2020b). The police, in their role as gatekeepers to the criminal justice system, have a significant role to play in improving these statistics (Barlow and Walklate, 2018). Their views about this project's potential to support investigations were an important part of this feasibility study.

The reasons for few cases being prosecuted are complex and wide-ranging (Bettinson and Robson, 2020). The Crown Prosecution Service (CPS) themselves have identified the need to secure more domestic abuse prosecutions, focusing on broader patterns of behaviour. This includes sourcing evidence such as emails, phone records, text messages and social media platforms which will provide records to be used as evidence in cases of coercive control (CPS, 2017). However, an increase in the collection of personal digital data has contributed to delays in investigative processes and a high dropout rate of victim/survivors. Victim/survivors expressed reluctance to share their data because they feel 'digitally strip searched' (HM Govt 2021, p8). Insight into the concerns, (or otherwise) of victims of domestic abuse about the role of AI in identifying and (risk) assessing perpetrators of domestic abuse is crucial. This includes their willingness to share their mobile phones with the police.

Aims and Objectives

This feasibility study ran from 13th December 2021 until the 30th March 2022. It aimed to test the suitability of Artificial Intelligence (specifically natural language processing methods as a tool to identify and risk assess (alleged) perpetrators of TFCC from digital communication held on mobile phones. Using secondary data from selected court transcripts the research aimed to:

1. establish the effectiveness (speed and accuracy) of using NLP and ML methods to analyse digital communication to identify perpetrators of domestic abuse.

2. establish the effectiveness (speed and accuracy) of using NLP and ML methods to analyse digital communication to identify indicators in the escalation of (alleged) abuse and risk.

3. see if the anonymity and discretion of using NLP methods for data collection encourages domestic abuse survivors to share digital data with police.

4. see if the police would embrace NLP methods as a tool to analyse digital data quickly and support with their investigations.

Overview of the research.

Ethical approval for this study was obtained from LSBU and Edgehill University's ethics committee. This research consists of four Work Packages namely:

Work Package 1: Survivor Survey

Two online surveys were conducted. The first aimed to capture victim/survivors' experiences of sharing digital data with the police. The team were keen to know if AI techniques would make victim/survivors of domestic abuse feel less 'digitally strip searched'.

Outcome: Evaluate survivor views on sharing their digital data and their attitudes relating to the tool created in work package 4.

Work Package 2: Police Survey

The second survey was with police staff and explored police attitudes regarding the model's potential to support police with their investigations and securing successful outcomes.

Outcome: Determine police officers' attitudes towards the tool created in work package 4.

Work Package 3: Digital Data Collection

Search criteria were established, and local and national media reports were scanned for the purpose of identifying domestic abuse cases that involved the use of digital communication. Once identified, relevant crown courts were contacted and permission was sought for extracts from the court case (eg CPS case, mitigation etc) that might contain digital communication as evidence (eg texts and emails). The subsequent transcriptions were then searched for examples of digital evidence and this data was then analysed as part of work package 4.

Outcome: Locate mobile phone data from online sources and court transcripts to be used for work package 4.

Work Package 4: Data Analysis and Modelling

The delivery of this work package took a two-pronged approach:

Mining Natural Language Sources: Text mining and natural language processing (NLP) techniques were used to automatically extract, analyse, summarise, and

assess the digital evidence obtained from the court transcripts. Innovative text analysis tools were developed to identify sent vs. replies in a message thread; identify the types of messages based on their content; identify coercive terms; quantify the sentiment and emotions expressed in messages; and cluster message threads in thematic groups to summarise the topics discussed. Messaging patterns such as the number, content types, submission times, length and duration of discussion, and frequency were also considered.

Modelling Perpetrator Behaviour: Machine Learning (ML) techniques were used to model the behaviour extracted from the "Mining Natural Language Sources" process. Behaviour extracted from court transcripts were considered as positive cases of coercive behaviour for training the ML model. Textual communication from other sources unrelated to the court records will be used as negative cases.

A lab-based evaluation was performed using the court transcript extracts to evaluate the NLP and ML models' performance (speed and accuracy) in identifying coercive behaviour from written text, and behaviour modelling. The outcome of this was to test the programme's ability to identify (alleged) perpetrators of domestic abuse and any indicators in the escalation of abuse and risk.

Outcome: Create model using data from work package 3 and evaluate model's performance.

Overview of Report Structure

The report will address each work package in turn. Every section will provide details of the study design and methodology, followed by the results and a discussion of the findings. Each section will end with a brief conclusion.

When each work package has been examined, the report will turn its attention to limitations associated with the research. The body of the report ends with a final conclusion of the whole report that brings all the strands together. It offers suggestions as to how this researched could be useful in informing or influencing policy relating to policing, domestic abuse and other forms of Violence Against Women and Girls (VAWG).

There then follows a full reference list and five appendices. The first provides some information about the authors' backgrounds, the remaining four provides additional/background information that relates to specifically WP4.

1. Victim/Survivor Survey.

1.1 Background Information

The abduction, rape, and murder of Sarah Everard by a serving police officer in March 2021 started a series of controversies that brought the issue of Violence Against Women and Girls (VAWG) into people's homes and placed it firmly on the political agenda (Ryan 2022). The murder of Sabina Nessa 18 months later fuelled the debate and highlighted the lack of priority that, it was argued, showed a level of disregard for Black and brown women (Mureithi, 2021). This concern was further highlighted when two police officers received prison sentences for misconduct in a public office because they took photos of two murdered sisters, Biba Henry and Nicole Smallman, and shared the images on WhatsApp groups. The victims' mother likened this behaviour to the photographing of 'lynching in the Deep South of the USA in order to highlight police racism and call for an urgent need to change police culture towards minoritized groups (Dodd, 2021). Investigations by the Independent Office for Police Conduct (2022) of inappropriate behaviour by police officers, predominantly from Charring Cross police station, identified further examples of misconduct and gross misconduct by officers who sent misogynistic and racist texts. These events led Maggie Blyth, Deputy Commissioner, and national police lead for VAWG, to acknowledge *"The last year has seen some tragic and shocking incidences of violence against women and girls. There have been abhorrent examples of abuse or misogyny by police officers"* (NPCC 2022).

Addressing police (mis)conduct and culture lay outside the aims of this study. Nevertheless, the context as described is incredibly pertinent as this research was conducted against this backdrop of controversies. It is very likely the above occurrences were very much at the forefront in the minds of those survivors/victims who participated in this survey. It is important to acknowledge this.

Prior to the above events, the Government was already working on its End-to-End Rape Review (HM Government 2021a). The report itself was published in June 2021. The review explored the arrest and prosecution of perpetrators of sexual violence and identified victim/survivors as feeling 'digitally strip searched' because of officers demands for seemingly irrelevant information such as search histories relating to shopping and holidays. Survivors also expressed concern at being left without their phones and access to vital support at a traumatic time and for long periods, with one victim/survivor reporting that the police kept her phone for four and a half years (Robbins, 2022).

The intrusive and seemingly irrelevant nature of the investigations coupled with the long waiting times required to analyse digital data means that survivors of domestic abuse often withdraw from domestic abuse cases before the matter is resolved (Robbins, 2022; HM Government, 2021a). AI technology can process digital data anonymously and faster than humans enabling anonymity of the victim/survivor and identifying perpetrators of domestic abuse quickly. This information was shared with victim/survivors as part of the on-line survey which formed Work Package 1.

1.2 Study Design & Method

The aim of this Work Package was to capture the attitudes of victim/survivors of domestic abuse towards police using AI as a tool to support their investigations. To our knowledge, this is the first time victim/survivors of domestic abuse have been

asked about the use of AI in the CJS. The survey was designed to assess whether survivor/victims would be more or less likely to share their digital data with the police if AI technology was used. This is because digital data (including that held on mobile phones) is extracted as a series of numbers with little/no human involvement and the anonymity afforded by this might influence victim/survivors' attitudes. Respondents were also invited to explain the reasons behind their expressed views. More specifically, the survey aimed to answer two research questions:

- Will the anonymity and discretion from the NLP techniques for data collection and analysis make victims of TFCC feel less 'digitally strip searched?
- Will this encourage them to share their digital data with the police?

Ethical approval was obtained from London South Bank University's ethics committee before the anonymous online survey was distributed. Participants were asked to score their responses to four questions on a Likert scale. There was also an opportunity for further comment via a free text box at the end of the survey. Demographic data (age, sex, ethnicity, and disability) was also collected. The survey went live on the 16th of February 2022 and closed on the 31st of March 2022.

Purposive sampling was used to obtain the views of victim/survivors. Survivor organisations known to members of the research team (including organisations supporting Black, Asian and minority ethnic survivors of domestic abuse) were approached for their help in recruiting victim/survivor respondents. This included promoting the research and distributing the survey link to survivors via their websites. Some organisations used the survey as a focus of discussion during survivor forums, on these occasions the PI offered to attend to answer any questions about the research. The PI' s attendance was at the discretion of the organisations.

Snowballing techniques were also used. The organisations mentioned above were asked to distribute a link to the survey amongst their wider networks. The link was also posted and promoted at regular intervals via social media networks namely Twitter and LinkedIn. As an incentive, victim/survivors of domestic abuse were offered the opportunity to enter a draw for the chance to win a £100 voucher.

1.3 Result

A total of 81 victim/survivors completed the survey. Three-quarters (74%) of participants were female, 22% male, with the remaining 4% preferring not to say. In terms of ethnicity, 78% of participants described themselves as White/Caucasian and 14% Black/African/Caribbean/Black British. One participant described herself as Asian and one participant was from a traveller community.

The survey was only open to participants aged 16 years and over. Ages were divided into categories beginning with 16-24 years (4% of participants) and then per decade (i.e., 25-34 years) thereafter until participants were asked if they were 75 years or over. Participants were represented across all age groups up to age 74. The survey was not completed by anyone aged 75 or over. The most represented age groups were victim/survivors aged between 35-44 years (27%) and 45-54 years (28%). Of the remaining participants 15% were aged 25-34, 19% were 55-64 and 6% were 65-74. One participant preferred not to say. This indicates that TFCC occurs throughout the lifespan and that this research is relevant to survivors of domestic abuse of all ages.

The survey showed that 70% of participants said they **would be likely or very likely to share information (for example text messages) stored on their mobile phones if the police used a computer programme to help them investigate domestic abuse**. 9% indicated they were unlikely to share this information, and fewer than 5% said they were very unlikely.

Remaining responses to the further three questions were more equally distributed. This suggests some ambiguity on behalf of the participants. 43% said they were **likely or very likely to worry about having their lifestyle scrutinised**, whilst 35% were unlikely or very unlikely. One fifth (20%) neither likely nor unlikely.

46% of participants were **likely or very likely to worry about their credibility as a victim/survivor of domestic abuse if their digital data were converted into a series of numbers and read by a computer,** with 36% indicating they were unlikely or very unlikely to be concerned about this. 14% were neither likely nor unlikely to worry.

58% of participants were **likely or very likely to worry that irrelevant information would be used against them in a police investigation.** Only 22% were neither likely nor unlikely to worry about this.

Although not the purpose of this research, the importance of technology in abusive relationships and the need for police to use this data as part of their investigations was clear in participants' responses provided in the free text box at the end of the survey. This included the mobile phone's ability to hold compelling evidence to secure a prosecution

'I begged the police to read through the very clever abusive techniques my ex-husband uses to abuse me via messages, but they've never done so' (Female White 35-44).

'My brother is dead due to CCB [coercive and controlling behaviour]. I know if his phone has been interrogated by police the chances of prosecution would have been much higher.' (Female White 35-44).

For one participant, AI offered validity to their accounts of abuse. They felt that AI would remove the emotions from their experiences and that their account would have more authority in the court room bringing with it some hope of justice.

'Sharing information from a mobile computer or phone feels as if it would take the emotions out of the situation which would be beneficial for me, as a survivor…the fact is data is recognised as being more legitimate in a court room than emotions (words that have been shown to be said can be nuanced less), would make me feel safer and happier about giving evidence.' (Female White 16-24).

A few participants explained that their hesitancy to embrace AI unconditionally related to how the data could be misused.

'I am very uncertain about my answers here. I am supportive of technologies that can identify abuse but at the same time worry about the data being mis-used.' (Female White 25-34)

'I would need to understand what data was used and how it might be conclusive to evidence before deciding if I found it likely to be harmful or useful. Female.' (White 35-44)

'Like all these things. It's not about the information it's about who has access and whether that access, changes for different reasons. So once agreeing the parameters change.' (Male Traveller 25-34)

"…where does big brother begin and end?" (Female White 45-54).

As outlined in the introduction, many coercive and controlling behaviours are complex and when taken in isolation can appear insignificant to the outsider. The need to understand digital data within a wider context was recognised by several participants who expressed concern about AI's ability to do this.

'Does the programme consider context? Obviously, some "in comments" between two individuals can seem innocent if you don't have a background of a relationship to set the scene as actual police officers would get.' (Female White 35-44)

'…. Would it [AI] be able to identify all aspects, the hidden passive content as well as the obvious and more outrageous content? If the software would be made so it could pick up on the aspects, that's great.' (Male White 35-44)

'….. does not take into account the tone of words, and phrases where would ordinarily look harmless but are used against you negatively. What looks like harmless "pet names" for example are used to degrade you. The AI system may not pick up the human element and like anything words are only understood in context. It may be likely serious cohesive control will be missed.' (Female White 25-34)

'So much intention and effect is specific to the circumstances. I feel this would be limited in use to only extremely clear-cut examples and very easy to misrepresent less clear-cut examples.' (Male White 35-44)

The survey findings also showed that seven of the victim/survivors of domestic abuse who commented in the survey shared a distrust of the Criminal Justice System.

'The criminal justice system doesn't support survivors of abuse, even when all the evidence is present - even if AI could gather the evidence, I believe it would be unlikely to increase conviction rates or offer protection as the justice system itself is in need of reform.' (Female Black/African/Caribbean/Black British 25-34)

However, some participants viewed AI technology as independent from the police albeit the lesser of two evils. Whilst cautious about its role, some participants seemed more willing to present their digital data for technological processing than human scrutiny suggesting there may be opportunities to use AI to engage victims more readily in the criminal justice process and build public trust.

'This [concern of data misuse] comes from a distrust of the police and family courts. I am in favour of this technology, but this doesn't mean I am not worried.' (Female White 25-34)

'I would rather my data [be] used than the police assess whether I was being abused or controlled. [area] police we're very helpful, I'm not sure how I'd feel if it was the MET [Metropolitan Police].' (Female Black/African/Caribbean/Black British 45-54)

'I have no faith that this system would work however it may be better than the current human model, which is open to corruption, misogyny, and abuse.' (Female White 64-74)

'If you are not tech savvy, I think I would have reservations about sharing other unrelated personal information on my phone.' (Female Black/African/Caribbean/Black British 45-54)

This scepticism was related to distrust in the police and concern as to how they would interpret digital data:

'…..I think a lot of male victims and survivors I talk to have a lot of mistrust. I think they would be worried about having the information used against them, having the 'why didn't you just leave' phrase thrown at them, I think they would like the idea of this, but would also have a lot la reservation too.' (Female White 45-54)

'Having worked with victims and survivors I feel many women would worry about their communication and how they have reacted to the years of abuse.' (Female White 45-54)

Victim/survivors in this survey were also worried about any bias inherent in the programme.

'I don't trust AI to be free of bias - it's only as good as the people that programme it. Already we're seeing AI being racist. By implication, this means that it's also sexist.' (Female White 55-64)

'I am worried about stereotypes that AI may generate related to survivors of domestic abuse.' (Female White 45-54)

Despite their reservations, participants could see the positives of using AI in police investigations. This included it's potential to make things easier for the victim/survivor.

'I think this takes the pressure off the woman to report and testify. I hope it gives a more accurate picture of how prevalent this is since repeat offences are often discounted.' (Female White 45-54)

The need for further understanding by victim/survivors, the wider population and the Criminal Justice System was highlighted. This included the need for further research.

'Survey is too simplistic to convey concerns but also hopes. So often victims of abuse are manipulated. There is a role of AI but much needs to be done to overcome concerns. Justice system is not yet sophisticated enough. Research important though.' (Female White 45-54)

'I think knowing exactly how the messages will be used and how the computer analyses them would be really important for people sharing the information. I know this is hard, because you need to build it first, but just an idea of how it will be analysed would encourage cooperation.' (Male White 24-35)

1.4 Discussion

The ability to secure over 80 responses in a little over six weeks, suggests that using AI to identify perpetrators of domestic abuse is important to victim/survivors. The themes outlined did not show any patterns in relation to specific demographics, but this is likely to be due to the small sample size. Although too small a sample to be generalisable to a wider population, the survey does suggest a curiosity by victim/survivors of domestic abuse who want to know more about how AI can be used in this way.

The survey indicated that survivor/victims of domestic abuse (70%) were willing to share their digital data with the police if AI programming was used. The outcome was less clear in relation to the victim/survivors' feelings of being digitally strip searched that saw responses distributed across the Likert scale. The free text offered some insight into this.

This survey findings showed concern about the programme's potential for bias. For some victim/survivors this was influenced by a mistrust of the police likely to be related to the poor policing practices outlined in section 1.1. Comments from other participants' however suggests that victim/survivors see AI as separate from the police.

Further qualitative research is required to offer an in-depth understanding of the victim/survivors concerns or otherwise of using AI as part of police investigations. This could include the potential benefits as well as providing further insights into their reservations about its use or misuse in this process. It would also be useful to see if and how using AI might circumvent some of the distrust victim/survivors have about the wider Criminal Justice System and if AI could be used to restore faith in the police and build public trust in the legislative process.

Speaking to Black and other minoritized victim/survivors is an important aspect of this as their experiences of abuse are shaped by their intersectional identities (Crenshaw 1991). These must be integrated into effective domestic abuse support offered to victim/survivors (Day & Gill 2020). This would also be consistent with the recommendations in the National Police Chief's Council report (NPCC, 2022).

1.5 Conclusion

The findings in this feasibility project reflect those from other research in that it suggests optimism about AI technology's role in decision making processes, with reservations about the police's ability to used it without bias (Aoki 2021, Hobson et al 2021, Kleinberg 2018).

The concern around bias is understandable given the limited public understanding of AI, its use and potential for misuse. The findings from this report echo that in Hobson et al's research (2021) who investigated public trust of decisions made by AI technology when compared to decisions made by the police. Results showed that participants supported the use of AI by police because it was deemed to be fair. Ludwig and Mullainathan's study (2021) shows that it is the data, not the algorithm that is open to bias. This suggests that educating the public about AI technology and its role in decision making could help build trust between the public and the police. Further research is required to understand this within a context of domestic abuse.

2. Police Survey

2.1 Background information

The Government has acknowledged that the Criminal Justice System (CJS) is failing victim/survivors of rape and sexual assault (HM Govt., 2021a). The Government also expressed concern about the backlog of domestic abuse cases in the courts because of the pandemic (HMICFRS, 2021). Despite the recent introduction of the Domestic Abuse Act, which commits to holding perpetrators responsible for their abuse, the Government (HM Govt., 2021a) has acknowledged that public trust and confidence in the CJS has eroded.

Recent reports and reviews from Government on VAWG and the CJS have shown that the volume of digital data and the length of time it takes to analyse it is a significant factor undermining police investigations and the prosecution process (HM Govt., 2021a; HM Govt., 2021b). Police forces report being overwhelmed by the exponential growth in the volume of digital evidence, with over 20,000 digital devices waiting to be processed. Chris Porter, the Met's director of forensic services, explained that "*The ability of policing to keep up with the demand is challenging*" (Robins, 2022).

Whilst using artificial intelligence (AI) within criminal justice settings is gaining momentum (Taylor, 2019; Lanier, 2019) it remains an under-researched area, specifically in the UK. Using Algorithms (a form of AI) to support decision-making within public sector and criminal justice agencies has a more established history. For example, Schoech et al. (1985) explored the potential of using AI to develop 'expert systems' (p 81) to support professional decision-making.

AI means different things to different people, making it difficult to define. In the field of computer science, it relates to data processing, learning, and evaluation phases (Zhou, 2017). In health and social care, the focus is on systems that can aid or assist to improve outcomes for service users. In the US, Asia, and Europe artificial intelligence now stands as a comparative model to traditional tools of decision making within criminal justice systems, which has relied on "humans" legislative frameworks (laws and policies) and intuition in decision making (Chiao 2019). Strong arguments exist against the use of technology such as AI because of the potential for biased data and the injustice that might exist because of this (Ryan, 2020). Concerns have also been raised over the accuracy of the data being inputted along with debates that criminal justice activities are complex and cannot be put into a number. Decisions that need to be made by professionals within the justice system are important[1]. The introduction of algorithms into criminal justice is increasingly perceived as flawed noting that algorithms fail because of substandard construction. (Kleinberg et.al., 2018) The positives however have been highlighted around the ability of machine learning to use large dataset for prediction (Sobjerg et.al., 2020).

AI has been used within criminal justice agencies to assist with decision making (Ludwig and Mullainathan, 2021). Machine learning (the training of an AI model on large datasets) offers a more efficient tool because algorithms are trained on large data sets allowing the extraction of more predictive signals reportedly making them

[1] whatworks-csc.org.uk 2020

more accurate. Machine learning is also suitable for more complex systems such as speech, text or video (Berk, 2018). This has resulted in the proliferation of algorithms across a wide range of criminal justice applications. AI technology is now able to undertake complex tasks that require cognitive capabilities such as making tacit judgements, sensing emotion, (Taylor, 2019) and driving processes (eg risk prediction) which had previously seemed impossible (Mahroof, 2019).

2.2 Study Design & Method

This aim of this work package was to determine the willingness of police staff to use AI as a tool in their investigations. A survey was therefore designed to answer the following research questions

- Will the police embrace NLP methods as a tool to analyse digital data quickly?
- Do the police think it will help the investigation process and how?

Ethical approval was obtained from London South Bank University's ethics committee before the anonymous online survey was distributed. Participants were asked to score their responses to four questions on a Likert scale. There were also two opportunities to comment, the first to elaborate on how AI could be useful and the second to share any other thoughts or comments about AI's potential or otherwise to assist with identifying suspects, investigating domestic abuse incidents, or influencing outcomes, such as charging a suspect. The survey went live on the 21st of February 2022 and closed on the 31st of March 2022.

Purposive sampling was used to obtain the views of police staff. Based on previous professional collaboration, the PI contacted Sophie Linden, Deputy Mayor for Policing and Crime at MOPAC asking for her support and introduction to relevant officers in the Metropolitan Police. Email correspondence and online meetings took place with to explain the purpose of the research. These staff were later sent a link to the online survey.

Snowballing techniques were also employed. Organisations, the individuals within it and academic contacts already known to the research team and who had built good relationships with police forces were also contacted and asked to distribute the link to their networks. The link was also posted and promoted at regular intervals via social media namely Twitter and LinkedIn.

2.3 Results

A total of 28 participants completed the survey. Eight were from the Metropolitan police, twelve from Thames Valley Police and six were from specialist units e.g., SEROCU, county lines, cybercrime. Two police staff did not disclose their location.

The majority (82%) of police staff who responded to the survey indicated that they would be **willing or very willing to use AI technology to support them in identifying domestic abuse**. Three-quarters felt that the technology **would be helpful in the investigation of domestic abuse incidents**.

75% of police respondents thought AI would be **helpful or very helpful in contributing to a successful outcome**, defined as charging a suspect in a

domestic abuse investigation. Almost 82% of participants thought AI would be **helpful or very helpful in building public confidence** in police investigations

Broadly speaking there was enthusiasm by the police about using AI in their investigations with some participants recognising it's potential in future policing.

'I think the benefits of this technology to support the active management of offenders in huge.' Police Officer

'we [cybercrime unit] use innovative technology to support traditional investigations. any assistance will support the investigator.' Police Officer

'Using AI for all digital investigations is clearly the future.' Police Officer

The significance of AI and its potential to assist police investigations because of its objectivity and ability to process data quickly, was seen as an asset by several in the survey.

'Reduce the amount of time an officer needs to spend analysing downloads, which is particularly important given the number of competing demands on officers.' Police Officer

'Technology like this can definitely speed up investigations.' Police Officer

'Because the amount of data held on a phone can be extensive, and when assessed by an officer the analysis is subjective and may be based upon their own preconceptions. Any 'objective' examination, that reduces officer subjectivity and workload could be hugely beneficial when understanding and assessing risk.' Police Officer

Some police staff shared their reservations about the programme's capacity to identify all relevant information within a wider context and like the results from the victim/survivor were unsure of its ability to do this without bias. The need for some 'back-up' was also recognised.

'Whilst this may flag pertinent messages, it begs the question as how/whether it is sufficiently reliable, both evidentially and for safeguarding purposes' Police Officer

'There are a number of caveats that have to be applied to AI-style analysis, including risks around false positives, or bias that is introduced by the human element that initially feeds the algorithm. Where these are appropriately mitigated, the technology would be "very helpful". If no mitigations are in place, the technology would be "very unhelpful"' Police Officer

'My concern with any form of AI is the fact that it can sometimes miss things that would indicate increased risk. Computers and AI applications can also fail, I would want some reassurance as to what back-up processes are in place.' Police Officer

Several participants suggested human involvement as a suitable safeguard to support AI and increase the reliability of analysis. Comments seemed positive, offering potential solutions to the issue rather than expressing concern per se.

'Like anything it's not a silver bullet as the accuracy would need to be verified by a human.' Police Officer

'Must have human oversight, with humans making decisions about how to proceed and act' Police Officer

'…. Clearly there needs to be a human factor in the decision making and review, but it can only but help.' Police Officer

Two participants recognised the potential for wider implications for the CJS, including the possible positive impact of AI on survivor engagement.

'…[if it] speeds up our processes, and reduces the possibility of victims disengaging then it would be worthwhile …' Police Officer

'Potential for lesser intrusion. Faster processing; supporting victim confidence and principles of Justice.?' Police Officer

But one participant urged a word of caution and recognised a tension amongst the public who may appreciate AI involvement but who might also be concerned about their data being used for other purposes. In do so s/he predicted one of the concerns raised in the survivor survey.

'Whilst the public may see the benefits of AI analysing the data, they may also have concerns about 'big brother' and what may be done with the data so communications would need to be carefully thought out.' Police Officer

Another participant identified the need for a wider buy-in across criminal justice agencies for AI to be useful.

'I think if used in criminal cases, Home Office and CPS support would be essential.' Police staff

2.4 Discussion

Like those in the victim/survivor questionnaire, results from the police survey showed unease about the programme's ability to understand the digital data within the wider context. The model's capacity to contextualise and classify information improves as greater volumes of data are provided. Embeddings (i.e., text conversion into computer readable format) that are utilised as part of text processing, notably BERT and GPT2 (see appendix 3), can retain the contextual information for each word or sentence. This allows any model trained on the embeddings to make inference based on contextualised information. As a feasibility study, the results have shown a minimum accuracy (Macro F-Score) of 88% (full details are available in section 4.5) which is higher than a random guess which is usually 50%. This is encouraging, although more data would be needed to improve the programme and provide the increased level of reassurance needed for police and victim/survivors regarding the programme's understanding of context.

Participants were also uneasy about the potential bias within the model. As outlined in Section 1.5, research shows that concern relating to AI bias is often misplaced

as it is the data that is subject to bias, not the programme itself. This is also true of this project, meaning that any bias that exists within this training data will reflect on the model outputs. Beyond the initial training, it is not possible for personnel to introduce bias to the programme. This means that analysis is entirely performed based on the model's training. For example, when there is new data to be analysed, the information is presented to the trained model to determine the presence or absence of abusive behaviour (Ludwig and Mullainathan, 2021)

Reducing the bias of the initial data was not within the remit of this feasibility study, though mitigation is possible through utilising larger sets of datum. This was not achievable at this stage of the research due to the limited number of court transcriptions and the short timeframe of the project. However further research over a more extended period, would enable this training to occur and the bias or otherwise of the programme could be better measured and understood.

Police participants suggested the need for human involvement to confirm the programme's analysis and make the ultimate decision eg proceeding with an investigation. The model is designed to be used to support decision-making rather than to replace it. Therefore, these valid concerns have already been taken into account during this project and will continue to inform later development.

The main advantages of using AI are the processing speed of the programme and the privacy it provides to survivors. Details of this are provided in WP4, which shows that the model can process large volumes of data quicker and make inference. For example, the preliminary experiments indicates that the programme can analyse a message of 224 words within 0.41 seconds and predict if it contains coercive behaviour. As outlined in WP 1 though tentative, survivors in this survey indicated a willingness to share their digital data knowing it will be processed by a machine rather than a police officer sifting through their personal information. It is encouraging that an AI approach could become a tool that supports investigative policing and victim engagement within criminal justice processes.

2.5 Conclusion

This survey indicates an appetite amongst the police for using AI technology to assist in their investigations. Whilst identified by one participant as 'the future', there were some reservations about its ability to understand digital communication within the wider context of coercive control. This is an important point since coercive control can often be misinterpreted, overlooked, and '*hidden in plain sight*' (Stark, 2007, p14).

The need for the police and the CJS to work together around the use of AI in identifying and risk assessing perpetrators of domestic abuse was highlighted in this survey and reflects the College of Policing's (2021) commitment to '*work with all parts of the policing landscape*' (p7) and strengthen partnerships across the CJS and the wider public sector.

It would be interesting to learn the reasons why and how police staff think that AI might be useful in building public trust. These ideas might then help forces to address the distrust of the police as expressed by some participants in the survivor survey. This is discussed in more detail in work package 4.

3. Digital Data Collection

The aim of this feasibility project was to test the viability of an Artificial Intelligence (AI) programme, specifically a Natural Language Processing (NLP) system, to identify perpetrators of domestic abuse using digital communication held on a mobile phone. Previous research has shown that machine learning models can be helpful to inform bail decisions in court (Berk et al., 2016) and in predicting risk based on criminal histories (Grogger et al., 2021). However, the data used in these studies is based on historical, 'fixed' data and does not consider the power dynamics of the relationship or the emotional responses of those involved.

3.1 Background information

This research used real life digital conversations between both proven and alleged perpetrators and victim/survivors of domestic abuse. Gaining access to these digital conversations was therefore central to this research.

The project ran from 13th December 2021 until 31st March 2022. Getting ethical approval from the relevant universities and obtaining the necessary information from criminal justice agencies (eg the police, CPS) or from the survivors directly was not realistic within these time constraints. Instead, the research team used court transcripts from domestic abuse cases where it was clear that technology was used as part of the abuse. Details of the ethical considerations are provided in Section 2.2

3.2 Study Design & Method

The research process began by establishing a list of inclusion and exclusion criteria that would be used to determine if a case was suitable for the purpose of this research. Cases were limited to incidents of domestic abuse within England and Wales only, as Scotland and Northern Ireland have separate criminal justice systems and criminal law. Cases were also restricted to those involving victims and perpetrators who were aged 16 years or over (in accordance with the Domestic Abuse Act 2021, s. 1 definition).

Searches for cases were conducted using manual Google searches and Google Alerts. The agreed search terms were "harassment", "coercive control", "messages", "bombarding" and suitable cases referenced in local and national newspapers were identified. News articles that made explicit reference to the use of technology for the purpose of the abuse with terms such as "harass", "inundate", "pester", "hound", "flood" and "assail" were used as they indicated a high volume of digital communication received by the victim/survivor.

The Law Pages (a publicly available legal resource and information website[2]) was also used to identify suitable cases. This website offers a comprehensive repository of information relating to court hearings. This includes (but is not limited to) case numbers, date and location of courts, and the name of the judge (which would later be required to authorize access to the transcript). The Law Pages also

[2] The Law Page website: https://www.thelawpages.com

indicated whether technology was used as part of the (alleged) offending in the details section of each case.

In all instances, cases that indicated not guilty pleas were prioritised. This is because details of the offence, including evidence of digital conversations are more likely to be presented to the court during a trial than when a defendant pleads guilty, and the court proceeds straight to the sentencing hearing.

When a suitable case was identified, using the case number, court and trial judge information provided by Law Pages, permission to access the transcript was requested from the trial judge. In the cases where an offender pleaded not guilty, the 'Evidence' and 'Opening of the Fact' were requested as it was anticipated that this would maximize the chances of finding the desired data. When an offender pleaded guilty, transcript requests were made in relation to the 'Opening of the Fact' and 'Mitigation' as it was anticipated that these would include any references to digital conversations during the court's hearings.

Once permission was granted by the presiding judge, requests for transcripts were made to the agencies authorised by the supervising court. A total of six transcription agencies were involved in this project and all data was received within twelve working days of the request.

Upon receipt of the transcript personal and identifying information of all parties was removed. Each transcript was processed manually to identify and extract all digital communication between the (alleged) perpetrator of domestic abuse and the victim/survivor. These records were then gathered in a spreadsheet and labelled '0' if the communication was from the (alleged) perpetrator and '1' if the communication was from the victim/survivor. The spreadsheet files were then ready for use as datasets for the NLP and data mining systems as outlined in work package 4.

Mindful of the time restraint and keen to 'train' the NLP system, whilst waiting for the court transcripts requests to be processed, preliminary data were also gathered from screenshots shared online by victim/survivors of domestic abuse. The search was conducted on Google using the same search terms identified for the court cases. The text on the screenshots was then manually entered into a spreadsheet and labelled '0' if the communication was from the alleged perpetrator and '1' if the communication was from the victim/survivor.

Ethical approval was obtained from the ethics committees of both London South Bank University and Edgehill University. This was to ensure that data shared between the two institutions met the necessary confidentiality standards and complied with the Data Protection Act 2018. Permission was obtained from the presiding judge at each of the selected courts. Names of both defendants and victim/survivors were not used in the preparation for the software. Transcripts were stored on a secure date repository that was accessed by a password protected laptop. No one was harmed in any way in the collection or analysis of data.

3.3 Results

The preliminary dataset built from digital conversation shared online includes 320 messages and 2,592 words.

The final dataset, built from court transcripts consisted of a total of 52 cases were identified across 28 courts. Of these, 30 transcript applications were made to the presiding judges. One application was rejected, and permission was received from a further 13 cases. Six transcripts were received (see section 5 for details) within the timescales of this project and used for analysis in Work Package 4. A total of 219 messages were extracted which consisted of 398 sentences and 3,087 words for this final dataset. This provided a usable dataset for the NLP and data mining systems explored and applied in this research.

3.4 Discussion

The decision to use court transcripts as data sources was a deliberate one and an attempt to circumvent the time constraints inherent in this project. Previous experience within the team had shown that it would not be possible to gain access to data from the police or Crown Prosecution Service (CPS) because of the complex and time-consuming process around accessing their data for research purposes. There was an element of risk associated with this approach since there was no guarantee that digital conversations would be considered in the court room and when this was the case, there was no certainty about the volume of data available.

Despite the limited time, this project identified 52 court cases where there was evidence of TFCC to support charges for domestic abuse related offences. Of the six transcriptions received from the courts and used in this project, all included digital conversations that were suitable for analysis in this study. This suggests that TFCC is a common factor in domestic abuse cases and highlights the need to identify this form of abuse quickly. Understanding the centrality of TFCC in the wider context of domestic abuse is therefore essential to assist criminal justice professionals with identifying and risk assessing alleged perpetrators of TFCC.

3.5 Conclusion

The purpose of this work package was to identify suitable data from real life digital conversations to test if it would enhance computational understanding of an NLP system. To our knowledge this is the first time NLP was used on a dataset built with this specific method of data collection – court transcripts. We did also include data collected in the same manner as other projects concerned with machine learning and NLP (Subramani, et al., 2019; Al-Garadi, et al., 2021). For this, we relied on digital conversations found online, which could have been modified or revised before posting, unlike the data collected from the court transcripts. The court data provides evidence collected by the police as part of their enquiries, which is then examined by the CPS before it is presented at court. In this way, it significantly increases the reliability of the raw data analysed in this research.

The data from the online searches and screenshots were used to' train' the NLP programme ahead of testing the system with extracts from court transcripts. Court transcripts, though limited in the quantity of data, did offer a viable option to acquire the necessary digital conversations for analysis in this research. All data collected were labelled and then employed in work package 4. The details of this stage of the research are considered below.

4. Digital Data Analysis and Modelling

Natural Language Processing (NLP) is a multi-disciplinary sub-field of Artificial Intelligence, Computer Science and Linguistics that concerns the methods and tools for analysing natural language text. NLP research resulted from a necessity to analyse text automatically and rapidly (Qin *et al.*, 2021). The evolution of computation technology and the increasing volume of readily accessible digital text has seen NLP further develop on tasks such as sentiment analysis, word sense disambiguation, multilingual summarisation and automatic term extraction (Ahmad and Brewster, Christopher Stevenson, 2017).

4.1 Background Information

This section focuses on research and development of NLP and Machine Learning (ML) tools to extract relevant information about the behaviour of perpetrators of TFCC and subsequently develop a model to automate decision making. Specifically, this section explores sentiment analysis and the classification of messages/digital communication according to the behaviour expressed (i.e., abusive or non-abusive). Different text representation approaches and how they affect model performance are also investigated. A list of core NLP tools relevant to this project and an explanation for the use of Natural Language Toolkit (NLTK) library are provided in Appendix 2.

A key component of analysing textual data is that of word embeddings. This term is used to describe how words are represented computationally. Words in a sentence are converted into a series of numbers called vectors and are determined by the embedding method used. Various methods exist to extract embeddings from text, the ones investigated and utilised in this study were Word2Vec (Mikolov *et al.*, 2013),: GloVe (Global Vectors for Word Representation) (Pennington, Socher and Manning, 2014), BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2019),: GPT2 (Generative Pre-trained Transformer 2) (Radford *et al.*, 2019) and ELMo: Embeddings from Language Model (Peters *et al.*, 2018). These are summarised in Appendix 3.

Machine learning (ML) classification is a process in which data instances are recognised and differentiated for better understanding. Single-label classification viewed as the conventional classification tool (Tsoumakas and Katakis, 2007; Dembczyński *et al.*, 2012) is used to define a process in which the instances are computationally categorised into only one of two or more classes. In this research binary classification (i.e. consists of two classes) were used, eg to determine the presence or absence of coercive behaviour in the digital communication.

Many machine learning classification methods exist. In this work package, we make use of three classification models, to show how various text representations can have an effect on the experimental data presented in *Table 1*. The classifiers are briefly presented in Appendix 4. Details of the method and metrics used to evaluate the classifiers in this work package are available in Appendix 5.

4.3 Study Design & Method

The aim of this work package was to developed tools to test the suitability of NLP methods in identifying and risk assessing (alleged) perpetrators of Technology Facilitated Coercive control (TFCC). Using digital communication extracted from

court transcripts of domestic abuse cases the aim of this work package was to answer the following research questions.

- Can NLP methods be used to analyse digital communication (eg texts, social media entries, emails etc) between victim and perpetrators to identify perpetrators of domestic abuse?
- Can NLP methods be used to analyse digital communication (eg texts, social media entries, emails etc) between victim and perpetrators to identify indicators in the escalation of (alleged) abuse and risk.
- If so, is this programme effective, in terms of speed and accuracy, in achieving these results?

This research used both NLP and ML techniques to achieve these aims. The study design and methods are presented in two subsections, namely: Mining Natural Language Processes and Modelling Perpetrator Behaviour.

4.3 Mining Natural Language Processes

Text mining and NLP techniques were used to automatically extract, analyse, summarise, and assess the digital communicating from the court transcripts as outlined in Work package 1. Text analysis tools were used to

- identify perpetrators' messages from communication threads.
- identify coercive terms.
- quantify the sentiment and emotions expressed in messages.
- identify messaging patterns such as message size.

Data was not available to allow identification of message frequency, submission times, duration of discussion.

The research team retrieved messages from the communication threads in the court transcripts only from the (alleged) perpetrators. This is because this research was focussed on understanding the behaviour of (alleged) perpetrators of TFCC. There was a total of 261 messages, from (alleged) perpetrators'. Eleven were removed because they were unlikely to provide information gain such and included texts include 'You are', 'you did', 'they', 'for what', 'you', 'at that', 'where are you?', 'who are you with?' and 'hun'. These were identified after applying the core NLP tools to the data e.g., stop-word remover will remove the message 'at that'. We enriched the modest data size with additional 242 perpetrators' messages obtained from various online repositories. Therefore, a total of 482 perpetrators' messages was used in this report

As noted earlier, single-label ML classification requires that data instances (i.e., messages) are labelled into only one of two classes (namely '0' for (alleged) perpetrator and '1' for victim/survivors). This is to train a model for future interpretation or inference. To determine the presence or absence of coercive behaviour within a text message, all 482 messages were labelled as having coercive behaviour. To obtain data instances that represents the absence of coercive behaviour, we retrieved conversational tweets that contains positive

sentiment via the Twitter application programming interface[3]. These were passed through NLTK sentiment analyser (see Appendix 2 for further details) to determine the polarity of the messages i.e., 'positive', 'negative' or 'neutral'. From these Tweets 530 messages of 'positive' polarity were randomly selected as data instances with absence of coercive behaviour. Thus, the experimental data consists of 1012 data instances (messages) representing coercive (48%) and non-coercive (52%) behaviour. The data characteristics is shown in **Error! Reference source not found.**.

Table 1: Experimental data characteristics

	Data Source	**Min message length**	**Max message length**	**Total messages**
Coercive	Online preliminary dataset	1	82	232
	Court transcripts (6 used)	1	224	250
Non-coercive	Twitter	3	36	530
Total messages				**1012**

4.4 Modelling perpetrator behaviour

ML techniques were used to model the behaviour extracted from the "Mining Natural Language Process". Specifically, the behaviour extracted from the digital communication were considered as positive cases while tweets from Twitter were used as negative cases. The behaviour (features) considered in the experiments includes only the embedding vectors (i.e., the integer values of the words and/or sentences within the messages). These were used to train and evaluate the performance of three ML models namely Random Forest, SVM Linear and RBF. We also considered four different text representation approaches i.e., BERT, GPT2, GloVe and Word2Vec as outlined in Appendix 3.

Lab-based evaluation was performed using k-fold cross validation on the experimental data to determine the performance of the 3 classifiers (see Appendix 5). Macro and Micro F-score (also known as Accuracy) were considered as performance metrics. The purpose was to determine the extent (in percentage) each classifier was able to identify the presence or absence of coercive behaviour from written text.

4.5 Results

This section contains the results from three classifiers used in this feasibility study namely Random Forest, SVM Linear and RBF (see Appendix 2). All were trained with embeddings from BERT, GPT2, GloVe and Word2Vec (See Appendix 3). Aggregate measures derived from confusion matrix, such as precision, recall, Micro and Macro F-Score (see appendix 4) were used to evaluate the predictive accuracy of the classifiers and the results are presented in Table 2. To evaluate the processing speed of the classifiers, we computed the time taken to train and evaluate each classifier model as shown in Table 2.

[3] https://developer.twitter.com/en/docs/twitter-api

In terms of prediction accuracy, the results show that GloVe embedding gave the best performance across all three classifiers with 100% Micro, Macro F-Score, precision and recall (highlighted in bold typeface in Table 2). Other embeddings namely, GPT2 and Word2vec also produced a 100% Micro and Macro F score with different classifiers.

Table 2: Experimental Results showing Macro and Micro F-Score performance of the classifiers with different embeddings

Embeddings & Classifiers		Macro F-Score	Micro F-Score	Precision	Recall	Training/Evaluation Time (seconds)
BERT	Random Forest	0.92	0.92	0.83	1.00	512
	Linear SVM	0.92	0.92	0.83	1.00	513
	RBF SVM	0.88	0.89	0.75	1.00	530
GPT2	Random Forest	1.00	1.00	1.00	1.00	213
	Linear SVM	1.00	1.00	1.00	1.00	217
	RBF SVM	0.96	0.96	0.92	1.00	218
GloVe	Random Forest	**1.00**	**1.00**	**1.00**	**1.00**	247
	Linear SVM	**1.00**	**1.00**	**1.00**	**1.00**	256
	RBF SVM	**1.00**	**1.00**	**1.00**	**1.00**	260
Word 2Vec	Random Forest	0.98	0.98	0.96	1.00	253
	Linear SVM	1.00	1.00	1.00	1.00	255
	RBF SVM	1.00	1.00	1.00	1.00	270

The classifier models also performed well on the experimental data with RBF SVM and BERT embedding producing the minimum Micro and Macro F-score of 88% and 89% respectively. These results fully satisfy objectives 1 and 2; and a part of objective 3 of this work package. Specifically, the embeddings (i.e., BERT, GPT2, GloVe and Word2vec) are capable of representing the contextual information in messages (objective 1) such that they can be used to indicate the degree/accuracy of abuse (e.g., Macro F-Score of 88%) to inform escalation decision (objective 2 and "accuracy" part of objective 3).

The processing time shown in *Table 1* satisfies the "speed" part of objective 3. Specifically, the minimum and maximum training/Evaluation time for the 1012 messages are 213 seconds and 530 seconds respectively. It is important to note that these values represent the overall time to train the classifier, and not the time taken to make a prediction on each individual message which should be significantly lower. In fact, our experiments using Linear SVM with GPT2 to predict messages of length 1 and 224 (i.e., minimum and maximum length in *Table 1*) produced 0.40 seconds and 0.41 seconds respectively. This shows the model's effectiveness in terms of speed (objective 3) in achieving the outcomes of objectives 1 and 2.

4.6 Discussion

It is important to put the experimental results into context. Generally, 100% performance is the desired outcome in any ML classification task, but this is very rare. When such performance is obtained, it is important to investigate the results further to ensure that they are generalisable. For example, 100% performance can be considered satisfactory with high confidence, If the task involves a situation that does not change over time and there are no exceptions, such as a ML model that computes the sum of binary numbers (e.g., 10 + 10). In this case, the outcome will

always be 20 and this value does not change, so 100% is acceptable. However, there are cases where the situations of a specific task are different such as the time taken to drive to the same place of work each day of the week. In this situation, the behaviour is a distribution rather than function which means that the outcome may be affected by many factors such as traffic, weather, time of the day etc. Thus, 100% is very unlikely because the nature of the situation is not a function.

ML tasks usually involve distribution that may overlap so large volumes of data are required to train models that approach full understanding of the situation. Given the modest data size used in this experiment (i.e., 1012 messages), the results are unlikely to generalise in real world scenario. That said, the evaluation method used (i.e., $k - fold$ cross validation) is appropriate in such situations and likely to produce the most objective view of the classifiers' performance (see Appendix 5).

4.7 Conclusion/relevance to policy

This feasibility project investigated abusive/non-abusive behaviour classification using three classifiers and four embeddings. Whilst work has been undertaken to identify abusive behaviour from text such as sexism and racism etc (Waseem and Hovy, 2016; Badjatiya *et al.*, 2017; Park and Fung, 2017; Rawat and Wang, 2017), to our knowledge this is the first time research has focussed on domestic abuse incidents that distinguishes between abusive and non-abusive behaviour.

5. Limitations

This project began on December 13th, 2021, and ended the 31st March 2022, a duration of just over three months. Ethical approval was received from LSBU on 8th of February 2022 allowing a little over seven weeks to collect data. These timescales proved challenging and were an integral part of the limitations associated with this feasibility study.

The surveys were distributed in a timely manner but could only be open to relevant participants for a limited time, namely six weeks for victim/survivors and five weeks for the police. Concerted attempts to engage directly with the identified groups and efforts to promote them on social media brought some modest success of the survivor survey, but the sample sizes remained small and cannot be generalisable, especially in relation to the police.

Different police forces have different approaches to engaging with research and these varying policies meant that central consent was required for some police areas but not others. Negotiating these bureaucracies was time consuming and often required a duplication in processes which meant that access to police officers was difficult and beyond the timescales of this project. A longer research period is necessary to gain larger participation from both victim/survivors and police, which would provide a greater insight into their attitudes towards using AI in investigations relating to domestic abuse.

Obtaining transcripts from courts was also more complex than anticipated. This is because of the complexity of both the court process and the payment systems at the lead university. Despite increasing pressures, court staff were extremely supportive of the team when trying to obtain court transcripts. Many had not themselves come across such a request before and thus were unclear of the process. Establishing the most efficient way to obtain the necessary data therefore involved some to-ing and fro-ing between the research team and court staff, which inevitably delayed the timing of the initial requests.

Once processed permission was then required from presiding judges who were not always forthcoming with their responses. This is understandable given the endemic capacity issues in court (Justice committee, 2022) which means that responding to such requests are not a judicial priority.

In addition, each court required the research team to use a transcription agency that had been vetted and approved by that court. Working out the different request and payment processes, though not difficult, was time consuming and created some unexpected delays with data collection.

All the transcription agencies required their fee in advance. The payment process within London South Bank University is designed to respond to invoice requests, something no transcription agency involved in this project was able or willing to do. Gaining information from LSBU about the best way to meet the needs of the transcription agencies and obtain the necessary internal approval was unclear and sometimes contradictory. Many of the transcription agencies were automatically deleted after 28 days because there had been no internal approval. This means the suppliers had to be set up on the system again and the process had to re-start. This too caused unanticipated setbacks and resulted in seven requested transcripts arriving after the end of the Project. These were unable to be included in the data analysis.

Empirical work by economists has for decades noted the limitations of AI, illustrating that in every data application, the data is incomplete, not fully representing either the objectives or the information that decision-makers possess. For example, professionals rely on much more information than is available to algorithms, and individual goals are often not well-represented by the outcomes provided to algorithms. (Cross 2020) However, a combination of machine learning and routine electronic information normally available at arraignment might be able to provide timely and useful domestic violence forecasts of risk. There are examples of successful forecasting in other criminal justice settings and for other kinds of crimes (Berk 2012). Moreover, machine learning forecasts can be delivered within a real time of several seconds. Although, a concern does remain as to whether the information routinely available electronically prior to an arraignment is sufficiently rich to produce usefully accurate forecasts.

The modest data size used in this feasibility study (i.e., 1012 messages), does raise issues of generalisability. To train a ML algorithm with a high level of confidence, large amounts of data are required. Unfortunately, this project was limited in respect of the short timeframe and availability of court data and consequently the data available for analysis and training of the ML algorithm was restricted. Without further opportunities to collect larger amounts of mobile phone communications between (alleged) domestic abusers and victim/survivors, the findings must be treated as insightful rather than conclusive. However, the results are useful in terms of establishing that the model demonstrates feasibility and further research should be undertaken to collect more data and generate a higher level of confidence in the AI tool.

6. Conclusion

This research resonates strongly with the current policy context with its use of machine learning and data modelling to analyse written communication between (alleged) perpetrators and victim/survivors. Text mining and NLP techniques have been used (sparingly) in the criminal justice arena, but no existing study has employed them to analyse TFCC in domestic abuse cases. This project differs from other research involving technology and domestic abuse in that it focuses on communication between perpetrators and victims at the time the (alleged) abuse occurs.

In addition, the research obtained the views of police staff and victim/survivors of domestic abuse about the use of AI in police investigations. To the team's knowledge, this is the first time, the views of victim/survivors of domestic abuse have been sought. This includes the opinions of Black and other minoritized victim/survivors which is consistent with the recommendations in the National Police Chief's Council report (NPCC, 2022). Early indications are encouraging, with both surveys indicating a cautious optimism about using AI in domestic abuse cases. When viewed within the wider literature there is some hope that educating the public about AI technology and its role in decision making could help influence policy, build trust between the public and the police and strengthen partnerships across the CJS and the wider public sector: an issue identified as important by the College of Policing.

Delays in the CJS, including the length of time police hold victim/survivors mobile phones, contributes to their decisions to withdraw from cases and, ultimately, means perpetrators are less likely to be brought to justice. The Government has awarded £5m to accelerate the capacity of police forces to acquire and manage

digital evidence and has instructed the College of Policing to ensure the police have the necessary skills, knowledge, and confidence to undertake effective and high-quality digital investigation (HM Govt. 2021a). This feasibility study offers a method that could quicken the police's ability to process digital data, cut down on the length of time they hold victim/survivor phones, limit delays in the CJS and reduce the number of victim/survivors of TFCC who withdraw from cases. There may also be scope to use AI to increase public confidence in the police. Further research is required to understand the range of possibilities open to the CJS because of AI and the best way of achieving these.

References

Ahmad, K. and Brewster, Christopher Stevenson, M., 2017, 'Words and Intelligence I', *SplringerLink - Text, Speech and Language Technology Book Series*, 35.

Al-Garadi, M.A., Kim, S., Guo, Y., Warren, E., Yang, Y.C., Lakamana, S. and Sarker, A., 2021. Natural Language Model for Automatic Identification of Intimate Partner Violence Reports from Twitter. *medRxiv*.

Aoki, N., 2021. The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior*, *114*, p.106572.

Badjatiya, P., Gupta, S., Gupta, M. and Varma, V., 2017, April. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion* (pp. 759-760).

Barlow, C. and Walklate, S., 2020. Policing intimate partner violence: The 'golden thread'of discretion. *Policing: a journal of policy and practice*, *14*(2), pp.404-413.

Berk, R.A., Sorenson, S.B. and Barnes, G., 2016. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, *13*(1), pp.94-115.

Bettinson, V. and Robson, J., 2020. Prosecuting Coercive Control: Reforming Storytelling in the Courtroom.

Chiao, V., 2019. Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. *International Journal of Law in Context*, *15*(2), pp.126-139.

Crenshaw, K., 1991. Race, gender, and sexual harassment. *s. Cal. l. Rev.*, *65*, p.1467.

Cross, T. (2020, June). An understanding of AI's limitations is starting to sink in. The Economist: Technology Quarterly. Available at https://www.economist.com/technology-quarterly/2020/06/11/an-understanding-of-ais-limitations-is-starting-to-sink-in

Day, A.S. and Gill, A.K., 2020. Applying intersectionality to partnerships between women's organizations and the criminal justice system in relation to domestic violence. *The British Journal of Criminology*, *60*(4), pp.830-850.

Dembczyński, K., Waegeman, W., Cheng, W. and Hüllermeier, E., 2012. On label dependence and loss minimization in multi-label classification. *Machine Learning*, *88*(1), pp.5-45.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dodd V., 2021, Met police officers plead guilty over photos taken at scene of sisters' deaths., The Guardian, Available at: https://www.theguardian.com/uk-news/2021/nov/02/met-police-officers-plead-guilty-over-photos-taken-at-scene-of-sisters-deaths?msclkid=5674e226c70f11ecab5295d2277d6f5b

Dragiewicz, M., Burgess, J., Matamoros-Fernández, A., Salter, M., Suzor, N.P., Woodlock, D. and Harris, B., 2018. Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, *18*(4), pp.609-625.

Elyezjy, N.T., 2015. Investigating crimes using text mining and network analysis.

Esposito, D. and Esposito, F., 2020. *Introducing Machine Learning*. Microsoft Press.

Felson, R.B. and Messner, S.F., 2000. The control motive in intimate partner violence. *Social psychology quarterly*, pp.86-94.

Grogger, J., Gupta, S., Ivandic, R. and Kirchmaier, T., 2021. Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases. *Journal of Empirical Legal Studies*, *18*(1), pp.90-130.

Hamberger, L.K., Larsen, S.E. and Lehrner, A., 2017. Coercive control in intimate partner violence. *Aggression and Violent Behavior*, *37*, pp.1-11.

Havard, T.E. and Lefevre, M., 2020. Beyond the power and control wheel: How abusive men manipulate mobile phone technologies to facilitate coercive control. *Journal of gender-based violence*, *4*(2), pp.223-239.

HM Government, 2021, End-to-End Rape Review Report on Findings and Actions HMSO. Available at: https://www.gov.uk/government/publications/end-to-end-rape-review-report-on-findings-and-actions [Accessed 8th April 2022].

HM Inspectorate of Constabulary Fire and Rescue Service (HMICFRS), 2021, Police Response to violence against women and Girls. Available at: https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/police-response-to-violence-against-women-and-girls-final-inspection-report.pdf [Accessed 31st March 2022.]

Hobson, Z., Yesberg, J.A., Bradford, B. and Jackson, J., 2021. Artificial fairness? Trust in algorithmic police decision-making. *Journal of experimental criminology*, pp.1-25.

IOPC, 2021, Failings identified in how the MPS handled missing persons reports for murdered sisters Independent Office for Police Conduct. Independent Office of Police Conduct 25th October., Available at: https://www.policeconduct.gov.uk/news/failings-identified-how-mps-handled-missing-persons-reports-murdered-sisters [Accessed 1st April 2022]

IOPC, 2022, Operation Hotton Learning report - January 2022, Independent Office of Police Conduct 25th October., Available at: https://www.policeconduct.gov.uk/sites/default/files/Operation%20Hotton%20Learning%20report%20-%20January%202022.pdf [Accessed 2nd April 2022]

Justice Committee, 2022, Court capacity: sixth report House of Commons, Available at: https://committees.parliament.uk/publications/21999/documents/163505/default/ [Accessed 27th April 2022].

Kleinberg, J. *et al.* 2018. Discrimination in the Age of Algorithms, *Journal of Legal Analysis*, 10(2005), pp. 113–174. doi: 10.1093/jla/laz001.

Lanier, P., Rodriguez, M., Verbiest, S., Bryant, K., Guan, T. and Zolotor, A., 2020. Preventing infant maltreatment with predictive analytics: Applying ethical principles to evidence-based child welfare policy. *Journal of family violence*, *35*(1), pp.1-13.

Ludwig, J. and Mullainathan, S., 2021. Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System. *Journal of Economic Perspectives*, *35*(4), pp.71-96.

Mahroof, K., 2019. A human-centric perspective exploring the readiness towards smart warehousing: The case of a large retail distribution warehouse. *International Journal of Information Management*, *45*, pp.176-190.

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mureithi A., 2010, Why isn't Sabina Nessa getting the attention Sarah Everard did?, open Democracy, Available at: https://www.opendemocracy.net/en/opendemocracyuk/why-isnt-sabina-nessa-getting-the-attention-sarah-everard-did/

National Police Chief's Council NPCC, 2022, Police progress against new framework on violence against women and girls, Available at: https://news.npcc.police.uk/releases/police-progress-against-new-framework-on-violence-against-women-and-girls [accessed 17th April 2022]

Stripe N., 2021, Domestic abuse and the criminal justice system, England and Wales: November 2020., Office for National Statistic (ONS), Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/domesticabuseandthecriminaljusticesystemenglandandwales/november2020

Pandey, R., 2020. *Text Mining for Social Harm and Criminal Justice Applications* (Doctoral dissertation).

Park, J.H. and Fung, P., 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

Peters, M. *et al.,* 2018, *Deep Contextualized Word Representations*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 2227–2237. doi:10.18653/v1/N18-1202.

Pilehvar, M.T. and Camacho-Collados, J., 2020. Embeddings in natural language processing: theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, *13*(4), pp.1-175.

Pinho, R. A., Brito, W. A. T., Motta, C. L. R., & Lima, P. V. 2017. Automatic crime report classification through a weightless neural network. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 165–170.

Qin, X., Liu, J., Wang, Y., Liu, Y., Deng, K., Ma, Y., Zou, K., Li, L. and Sun, X., 2021. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *Journal of Clinical Epidemiology*, *133*, pp.121-129.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), p.9.

Rawat, W. and Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, *29*(9), pp.2352-2449.

Riya, R., & Gandotra, N., 2016. Text Mining on Criminal Documents. Int J Advances in Electronics and Computer Science, 3(9), 28–31.

Robbins J., 2022. Police have digital backlog of more than 20,000 devices waiting to be examined., The Justice Gap. Available at: https://www.thejusticegap.com/police-have-digital-backlog-of-than-20000-devices-waiting-to-be-examined/ [Accessed 2nd April 2022].

Ryan, M., 2020. In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), pp.2749-2767.

Schoech, D., Jennings, H., Schkade, L.L. and Hooper-Russell, C., 1985. Expert systems: Artificial intelligence for professional decisions. Computers in Human Services, 1(1), pp.81-115.

Søbjerg, L.M., Taylor, B.J., Przeperski, J., Horvat, S., Nouman, H. and Harvey, D., 2021. Using risk factor statistics in decision-making: prospects and challenges. *European Journal of Social Work*, *24*(5), pp.788-801.

Stark, E., 2007. Coercive control: the entrapment of women in personal life Oxford.

Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y. and Shakeel, H., 2019. Deep learning for multi-class identification from domestic violence online posts. *IEEE Access*, *7*, pp.46210-46224.

Taylor, L., 2019. AI may be a solution to the social care crisis, but what are the legal concerns? Available at: https://www.computerweekly.com/opinion/AI-may-be-a-solution-to-the-social-care-crisis-but-what-are-the-legal-concerns

Tsoumakas, G. and Katakis, I., 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, *3*(3), pp.1-13.

Waseem, Z. and Hovy, D., 2016, June. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

Wiener, C., 2017. Seeing what is 'invisible in plain sight': Policing coercive control. *The Howard Journal of Crime and Justice*, *56*(4), pp.500-515.

Williamson, E., 2010. Living in the world of the domestic violence perpetrator: negotiating the unreality of coercive control, *Violence Against Women*, 16(12), pp. 1412–1423.

Woodlock, D., McKenzie, M., Western, D. and Harris, B., 2020. Technology as a weapon in domestic violence: Responding to digital coercive control. *Australian social work*, *73*(3), pp.368-380.

Women's Aid, 2019. What is coercive control? Available at: https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/coercive-control/ [Accessed 31st March 2022.]

Xu, J., 2021. Research on Judicial Big Data Text Mining and Sentencing Prediction Model. In *Journal of Physics: Conference Series* (Vol. 1883, No. 1, p. 012158). IOP Publishing.

Zhou, J., Witt, K., Cao, X., Chen, C. and Wang, X., 2017. Predicting reoffending using the Structured Assessment of Violence Risk in Youth (SAVRY): A 5-year follow-up study of male juvenile offenders in Hunan Province, China. *PLoS one*, *12*(1), p.e0169251.

Appendix 1

About the authors

Dr Tirion Havard is an Associate Professor in the Social Work department at London South Bank University. She draws on her experiences as a probation officer working with perpetrators of domestic abuse and gang members to inform her research with women survivors of abuse. She is Principal Investigator for a British Academy/Nuffield Trust project about reintegrating women with convictions into their local communities, evaluating a Serious Violence Intervention and Prevention project at Kent County Council and the use of Independent Domestic Violence Advocates (IDVA) in Surrey hospitals. She is also co-investigator for Project vigilant, Thames Valley police's strategy to tackle sexual harassment in the night-time economy. Tirion has also worked with the London Borough of Southwark, The Mayor's Office London and English Parliament advising them on coercive control and how it relates to domestic abuse and young women and girls in gangs

Dr Nonso Nnamoko is a lecturer in Computer Science at Edge Hill University (EHU). He holds a PhD in Artificial Intelligence and has research experience in the areas of Machine Learning, Big Data analytics, Software Engineering, Natural Language Processing and Computational Linguistics. Nonso has been involved in many national and international research projects including two European Union Horizon 2020 projects CROSSMINER and TYPHON (each awarded approximately €4.5 M). He is also involved in a recent project awarded £120K by the British Academy and Nuffield Foundation. Nonso currently leads two academic start-up projects funded by Innovate UK, namely projects CyberSignature and LimbProve. He has authored many articles in reputable interdisciplinary journals, books and conferences.

Dr Chris Magill is a Senior Lecturer in Law with Criminology at the University of Brighton. For 10 years prior to academia, she was a Principal Social Researcher for Government departments, including the Home Office, Ministry of Justice and the Crown Prosecution Service. She is currently a Principal Investigator on an evaluation of Project Vigilant, an international award-winning perpetrator-focused initiative to prevent sexual offending in the night-time economy and Co-Investigator on several projects focusing on violence against women and girls, including a British Academy/Nuffield Trust project and the use of IDVAs in Surrey hospitals. Chris is also a Trustee for RISE, a Sussex-based charity that supports people affected by domestic abuse and violence.

Cyndie Demeocq got her MSc from the school of engineering at the London South Bank University. She is especially interested in the development and usage of multimodal learning and multimodal datasets to address criminal and abusive behaviors.

Jack Joseph Procter is a programmer with an affinity for software and games development, history, linguistics and literature. He studied computer science at Edge Hill University, receiving a study award for his dissertation piece. Currently researches machine and reinforcement learning, natural language processing and neural networks

Denise Harvey is a qualified social Worker with over 23 years' experience working with children and families services. She has over 17 years of practice experience within Youth Justice and frontline children service as both a practitioner and senior/Operational Manager. She has extensive experience of working with high-risk offenders as well as managing high profile government projects that rolled out new ways of working within Youth Justice, and Social Care such as Reclaiming Social Work and FIP. Denise is currently a Senior Lecturer in Social Work and Consultant helping small businesses in setting up Social Care provisions. Denise has been a key member of the research team on a British Academy funded study of

professional decision-making in child protection, the Seeing Through the Eyes of Experienced Practitioners (STEEP) project. Denise was the Principal Research Advisor and an Expert Panel Member, based upon her extensive experience in social work practice.

Professor Vanessa Bettinson is a professor of Criminal Law and Criminal Justice. She is known as an international expert for her work on domestic violence and abuse and has written extensively on the issue of criminalising coercive control. Vanessa is currently engaged with colleagues from Warwick and Leicester Universities in empirical research exploring criminal justice professionals' understandings of coercive and controlling behaviour. She provided expert support to the defence counsel in the case of Sally Challen and is now working with academic colleagues on a variety of issues relating to the implementation of coercive control offences and the creation of new defences.

Appendix 2

Core Natural Language Processing Tools

• Tokeniser: Also known as lexical analyser, is designed to split a sentence/ string into tokens. Tokens are a set of characters that have a meaning by themselves (Jackson and Moulinier, 2002). A simple tokeniser splits a string by white space, but a more efficient tokeniser can use other techniques to separate elements eg punctuation and abbreviations (Loper and Bird, 2002).

• Part-of-Speech (PoS) Tagger: This tool tags every token (eg noun, verb, preposition) with its own PoS. It can be based on rules or trained on large annotated datasets to determine probabilities (Loper and Bird, 2002).

• Lemmatiser:lidentifies lemmas, i.e. the dictionary forms, of words and are frequently based on rules and dictionaries (Manning, Christopher D Raghavan, Prabhakar Schütze, 2008).

• Stemmer: It is a tool that uses heuristics to delete prefixes and suffixes of words in order to find its root. Stemmers are faster than a lemmatisers, however they are less accurate, especially in highly inflected languages, such as Spanish or French (Manning, Christopher D Raghavan, Prabhakar Schütze, 2008).

- Stop-word remover: It is a tool utilised to remove semantically irrelevant words, often words which serve to connect elements of a sentence grammatically. Through removing these words, that contain little to no information, the remaining textual data can be further processed. In the literature we can find multiple libraries that offer this kind of tools, such as Freeling (Padró and Stanilovsky, 2012), Stanford CoreNLP (Manning *et al.*, 2014) and Natural Language Toolkit (NLTK) (Loper and Bird, 2002). We have decided to use NLTK for the following reasons:

1. **Accessibility** - NLTK is an open-source project, hosting thorough and readily accessible information regarding its functionality. This greatly aids in implementation and troubleshooting, as the library has received continued support since its inception. The transparency afforded by this structure, permits configuration of tools available within the library. This inherent flexibility affords many avenues to optimise the work package, pre-processing data in a manner most efficient for the experiments.
2. **Versatility** - NLTK offers a myriad of tools, possessing functionalities for many applications essential to Natural Language Processing. As such its implementation into the work package is kept concise and efficient as all elements of pre-processing, as described prior, can be conducted with NLTK's functionality. These include Tokenisation, PoS Tagging, Lemmatisation and Stemming, each of which offer a large degree of scalability.

Appendix 3

Embeddings

Word2Vec is a technique utilised within NLP to extract information from text, by converting it into a numerical representation. The process of obtaining word embeddings is conducted by training a model with a large amount of data, from which it can learn the meanings of words based on their linguistic relation to one another (Mikolov *et al.*, 2013). There exist two proposed methods that operate under Word2Vec's methodology, that of Continuous Bag-of-Words and Skip-Gram. The former predicts a word's meaning based on the context of the sequence, by summing the vectors of each few words adjacent to the current, it then predicts this word by comparing against the sum. Skip-Gram operates inversely, by utilising the current word to predict those adjacent on either side. For this work package a CBOW Word2Vec model is utilised.

GloVe is an unsupervised machine learning algorithm, that obtains vectors from text supplied during its creation and subsequent training. It is from this textual data that it assigns vectors to each instance of a word, these are then grouped based on their similarity and relation to one another. The distance between vectors can be used to determine a given word's relation, in terms of linguistics, to others present within the model. Sequences of words that are presented to the model can be assigned an appropriate vector, based on this grouping and distance process. Words that share a high degree of similarity are listed together and can be used to determine the relation between new sequences. GloVe offers a variety of pre-trained models, ranging in size and subject matter. In this work package, GloVe's 'Wikipedia 2014, 300 dimensions' model is utilised, having been trained on a generalised and large corpus of data (Pennington, Socher and Manning, 2014).

BERT is a pretrained model developed by Google. Its intended uses are for 'next sentence prediction' (tasked with predicting the feasibility of a subsequent sentence, or omission of the first) and the prediction of vectors based on the context of each word. The result of this pre-training was BERT's ability to learn contextual embeddings for words, this inherent benefit of BERT's structure ensures that contextual data is maintained. During the processing of text, each word references those adjacent to contextualise itself. For this work package, the 'Regular' model is utilised as it offers a balance of embedding information and timeliness (Devlin *et al.*, 2019).

GPT2 is an open-source artificial intelligence, developed by OpenAI (Radford *et al.*, 2019), that is generalised in its intended functions. GPT2 is capable of translating, summarising and generating text based on input data. A benefit of its open and generalised architecture is the ability to generate embeddings for sequences. This flexibility allows for the generation of embeddings, even if presented with novel words, and as such gives the model a degree of adaptability.

Appendix 4

Machine Learning Classification

There exists a large variety of methods that address the problem using different approaches. In fact, text classification algorithms are at the heart of a variety of software systems that process text data at scale. For example, e-mail software uses text classification to determine whether incoming mail is sent to the inbox or filtered into the spam folder. Discussion forums use text classification to determine whether comments should be flagged as inappropriate (Sapienza *et al.*, 2018; Gharibshah, J. Papalexakis, E. E. Faloutsos, 2020). Another common type of text classification is sentiment analysis, whose goal is to identify the polarity of text content i.e., the type of opinion expressed in the text. This can take the form of a binary like/dislike rating, or a more granular set of options, such as a star rating from 1 to 5. Examples of sentiment analysis include analysing Twitter posts to determine if people liked/disliked a particular movie or extrapolating the general public's opinion of a new prime minister. In this work package, we make use of three classification models, to show how various text representation can have an effect on the experimental data described in *Table 1*. The classifiers are briefly presented below.

Support Vector Machine (SVM)

A Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is a machine learning algorithm designed for binary classification problems. However, it can be used in multi-class tasks using a one-vs-rest strategy. The idea behind SVMs is to map vectors from a training data set, into a very high dimensional space, i.e., a hyper-space. Then, the optimal hyperplane is identified that can separate the vectors according to their class. New examples, such as those found in a testing data set, are mapped into the same hyper-space to predict their class based on which side of the hyper-plane they fall into. The mapping of vectors is done using linear algebra operations through the application of kernels such as the Radial Basis Function (RBF), the linear or the polynomial. Due to its robustness and relative simplicity, SVMs have become a classic method in the literature for performing single-label classification tasks (Pisner and Schnyer, 2020). Furthermore, SVMs supports both dense or sparse vector representations in line with our requirements.

In this work package, the experiments based on SVM are done using LibSVM (Chang and Lin, 2011). LibSVM, is a library that has been ported to different languages, like Java, Python and Perl. It is multi-thread and capable of implementing automatically a one-vs-rest strategy for multi-class problems. More importantly, in this work package we make use of both linear and RBF SVM. In other words, we explored two different SVM flavours: one that uses an RBF kernel and another using kernel for mapping the vectors into a high dimensional space.

Random Forest

Random Forest algorithm (Breiman, 2001) is an ensemble learning algorithm, that has applications in classification and regression tasks. Structurally, a Random Forest acts as an aggregator of predictions, as within the Forest are a set number of Decision Trees. Decision Trees (Quinlan, 1986) consist of numerous logic checks, each analogous to the leaves of a tree, with the path between each called a branch. The algorithm operates by receiving data, and then constructing a series of branches that extend and split at leaves. Upon construction each branch within the tree is traversed, and upon reaching a leaf, a decision is made based on the attributes in question. Leaves that have been traversed are assigned a degree (value) of confidence. At each leaf presented, a new branch can begin, further splitting at subsequent leaves. The aforementioned process repeats until all branches have been exhausted. The

branches are evaluated, and the instance with the highest confidence is selected as the prediction, from which a label is assigned. The outcomes of each of these Trees are then compared and a consensus is reached. Whichever prediction received a majority within the Forest, is selected as the final classification. The inherent advantage of this method is the mitigation of anomalous or low confidence predictions. The data is iterated through by each Tree within the Forest, with their predictions made in isolation, and as such a more confident prediction can be achieved via this algorithm.

Appendix 5

In this work package, three classification models were used to show how various text representation can have an effect on the experimental data described in *Table 1*.

Classifier Evaluation Methods and Metrics

In this section, we present the. As evaluation method, stratified k-fold ($k = 10$) cross-validation is commonly used where there is data paucity which is the case with the experiments presented in this work package (James *et al.*, 2013). In this evaluation method, the training data is randomly partitioned into 10 equal size subsets, taking the class distribution into account. During training, one of the k subsets is retained as the validation data, and the remaining $k - 1$ subsets are used as training data. The process is repeated k times, with each of the k subsets used exactly once as the validation data. The k results from the folds are then combined to produce a single result.

Single-label classifiers are commonly evaluated using metrics based on confusion matrices. We present in the following paragraphs, a brief description of each evaluation metric used. To understand the metrics better, we present, in Table 3, the structure of a general confusion matrix from which all metrics are deduced. Although the example matrix shows two classes, it can be extended to accommodate as many classes as necessary, depending on the dataset being evaluated. In Table 3, I_{nm} denotes the number of instances that belong to class *n* and were predicted as class *m*. A confusion matrix with zero off-diagonal values corresponds to the evaluation of a method that performs ideally, without any classification errors.

Table 3: General structure of the confusion matrix

		Predicted Condition	
		$Class_1$	$Class_2$
True Condition	$Class_1$	I_{11}	I_{21}
	$Class_2$	I_{12}	I_{22}

Precision of a single-label classifier for a determined class n is defined in Equation 1

$$Precision(n) = \frac{I_{nn}}{\sum_{m=0}^{c} I_{nm}} \quad (1)$$

where I_{nn} is the number of instances that were correctly predicted for class n, c is the total number of classes present in the dataset and I_{nm} is the number of instances of class n that were predicted as class m.

The **Recall** of a class n in a single-label classifier is defined in Equation 2

$$Recall(n) = \frac{I_{nn}}{\sum_{m=0}^{c} I_{mn}} \quad (2)$$

where I_{nn} is the number of instances that were correctly predicted for class n, c is the total number of classes present in the data set and I_{mn} is the number of instances of class m that were predicted as class n.

The **F-Score or F-1** is the harmonic mean of Precision and Recall. We present its definition in Equation 3

$$FScore(n) = 2 \cdot \frac{Precision(n) \times Recall(n)}{Precision(n) + Recall(n)} \quad (3)$$

In the literature, we can also find Macro and Micro versions of F-Score. The former is the averaged F-score. More specifically, we present its definition in Equation 4

$$Macro\ FScore = \frac{\sum_{n=0}^{c} FScore(n)}{c} \quad (4)$$

where c is the number of classes present in the data set. The Macro F-Score indicates how well the classifier performs on all classes regardless of the class size.

As shown in Equation 5, Micro F-Score is a weighted average, i.e. considers the proportion of each class in the data set. It should be noted that the Micro F-score is also known as Accuracy.

$$Micro\ FScore = \frac{\sum_{n=0}^{c} I_{nn}}{\sum_{n=0}^{c} \sum_{n=0}^{c} I_{nm}} \quad (5)$$