The Use of Computational Text Mining Methods to Detect and Understand Domestic Abuse

Jessica Lilly Neubauer

A dissertation submitted in partial fulfilment of the requirements for the degree of

Master of Philosophy

in Cybersecurity

University College London

2023

Please be aware that the following work contains descriptions of violence and abuse that some readers may find distressing.

Declaration

I, Jessica Lilly Neubauer, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed _______________

Date ____ / ____ /_____

Acknowledgements

I'd like to thank both my supervisors, Dr. Leonie Maria Tanczer and Dr. Enrico Mariconti, for their support, encouragement and understanding over the two years we were working together.
I'm also grateful for the support and laughter provided by the members of the Gender and Tech research group – Demelza, Niamh, Izzy, Kyle, Francesca, Maddy and Nikos.
Thank you to Lifang Li, Meghan Knittel and Demelza Luna Reaver who gave their time, attention and expertise for labelling the dataset of psychologically abusive behaviours.
Thank you to the members of the VISION Consortium at City University, I benefitted hugely from their valuable advice and experience in Domestic Abuse research.
Thank you to the staff and officers at Warwickshire Police who created such a warm and engaging atmosphere when they hosted me for a research visit.
Thank you to Mark Warner for advice and feedback on papers.
Thank you to Maria Schett for mentoring, support, and some very interesting chats.
Thanks to my friends, family and flatmates. Particularly my parents, my brothers, Karin, Marco, Thessa, Tina, Ola, Dorina, Phoebe, Josh, Robyn, Joanna, Shahid, Zoe, Erin, David, Rabia, Rose, Nicole & Katie for listening to me moan a lot and providing much needed perspective. Thanks especially to Lili and Milly who had the inside view on what it's really like to be a research student.
Thanks to Professor Jane Waldfogel for providing much needed insight into the academic world and her valuable advice and encouragement over the years.
Thanks to the other students in the Cybersecurity CDT – Aliai, Charlie, Demelza, Jay, Filippo, Marilyne, Karolina and Nadine for being intelligent and inspiring colleagues.
And to other PhD students in Crime Science and STEaPP who made Shropshire House such a fun place to work and study.
Thanks to all the staff in the UCL Computer Science department which was my professional home for five years, particularly to Nicolas Gold for providing inspiring and supportive supervision and mentoring, and giving me the opportunity to lecture.
Thanks to the members of the UCL Mindfulness Society who were wise, creative and empathetic friends for the past five years.
Thanks to all my musical and creative collaborators, particularly Dave, Debs, Gian, Erin and Ola, who kept me (sort of) sane whilst trying to balance research and music.
Thanks to my examiners, Kate Bowers and Lisa Tompson, for making the viva process such an enjoyable experience!

Abstract

Domestic Abuse (DA) is a widespread problem which causes major harm to victim-survivors. Psychological abuse is a common form of DA, and has a significant negative impact on victims, but is still not well defined or understood. Existing survey-based methodologies for researching psychological abuse could be complimented by computational social science methodologies using social media data. This thesis discusses the use of computational text mining in DA research, and seeks to contribute to this work through the creation of a dataset and a machine learning classifier to identify types of psychological abuse.

A systematic literature review was conducted to give an overview of current work applying text mining methodologies in the study of DA and identified a gap in the literature regarding automatic identification of psychologically abusive behaviour.

A dataset (n=2000) of Reddit posts was developed and labelled using an annotation scheme of six types of psychologically abusive behaviour. The annotation scheme was developed based on a literature review of existing measures and frameworks for psychological abuse, and refined over a series of expert discussions.

Finally, a variety of machine learning classification models were trained on the dataset of psychologically abusive behaviours. A DistilBERT pre-trained model performed well (F1-score = 0.81) at classifying *Threatening, Intimidating and Punishing* behaviour. However, machine learning models were not successful at classifying other types of psychological abuse, due to the small size of the dataset and highly imbalanced classes.

The thesis demonstrates that computational text analysis tools are useful for analysing large amounts of text data about DA, and providing insights into experiences of abuse that go beyond traditional qualitative methods. However, the thesis also illustrates the limitations of computational methods, which struggle to work well in the context of wider disagreements and debates about what constitutes abuse.

Impact Statement

This thesis aims to use computational methods in a field with a high potential for real-world impact: domestic abuse (DA) research. The findings of the thesis have the potential for impact in both research and practice across several fields.

Many social scientists, including those studying DA, are keen to exploit modern computational methods to make their work more efficient and open up new sources of data. This thesis offers a starting point for DA scholars to use similar methods in their own work, and has deliberately phrased descriptions of computational methods to make them more accessible to a non-technical audience.

The systematic literature review described in this thesis is the first of its kind to examine computational text analysis methods in intimate partner violence research. This provides a foundation for researchers wanting to use similar methods, and contributes to the development of the field by encouraging and facilitating the use of new, innovative research methods. Furthermore, the systematic literature review offers DA advocacy and support sector organisations a starting point to understand the potential for computational methods and how these might facilitate research with the data that they already hold.

The thesis also represents a contribution to research about psychological abuse. A new annotation scheme built on existing measures and frameworks of psychological abuse to develop a new typology and further extend this into a detailed tool for labelling machine learning datasets. This six-label categorisation could be used by researchers wishing to conceptualise psychological abuse for their own projects or datasets, and could also be useful for practitioners working with victim-survivors of psychological abuse.

For computer scientists and data scientists, the thesis offers insight into using computational methods in a social science context. It offers lessons and best practice for labelling new machine learning datasets, especially in topics where humans find it difficult to agree. This could lead to impact within the field of Responsible Artificial Intelligence, by contributing to the discussion about 'data explainability' and providing an example of how to make a dataset fully transparent and 'explainable'.

As part of a research group focusing on technology-facilitated abuse, the information in this thesis has been presented by the author at regular informal and formal meetings with DA advocacy organisations (such as Refuge) and key actors (such as the National Police Chiefs Council and the Metropolitan Police) in the DA sector. The author has also been a member of the VISION consortium at City University and has contributed to workshops and discussions about the use of Natural Language Processing in research about DA. Findings from this thesis were also discussed with members of a UK police force during a week-long research visit in August 2022.

The systematic literature review presented within this thesis has been published in the Journal of Family Violence and will be presented as part of a panel at the European Conference of Domestic Violence in Reykjavik, Iceland in September 2023.

Funding declaration

This research was supported by a Engineering and Physical Sciences Research Council studentship through the UCL Centre for Doctoral Training in Cybersecurity (Grant number EP/S022503/1)

Table of Contents

List of Tables

List of Figures

Glossary

Term	Definition
Accuracy	Evaluation metric which measures the percentage of instances to which a model assigns the correct label. It is calculated by dividing the total number of correct predictions of a model over the total number of predictions the model makes.
Adam	An optimisation algorithm for deep learning networks which helps the model converge to the optimal solution (Zhang, 2018).
Annotation	(Also called *label*) A class or category assigned to an instance in a dataset. (e.g. “contains abuse”, “does not contain abuse”)
Balanced Dataset	A dataset which contains relatively equal numbers of instances from each ‘class’ (category).
BERT	A large pre-trained language model trained on a very large corpus of data taken from the internet.
Bi-gram	A combination of two words which appear together in a text (are *collocated*).
Class	A category or label used in a dataset (e.g. “contains abuse”, “does not contain abuse”).
Classification	In Machine Learning, classification tasks are concerned with categorising and input into one or more output classes.
Classifier	A Machine Learning Classifier is a model which takes data instances as input and predicts which class each instance should fall into, given the data it’s been trained on.

Clustering Algorithm	An unsupervised learning algorithm that can uncover related clusters of information in a dataset without needing the dataset to be labelled/annotated.
Collocations	Refers to words that appear together or near to each other in a text.
Computational Text Mining/Analysis	A set of techniques which use algorithms to understand, categorise or extract information from unstructured text data.
Convolutional Neural Network (CNN)	A type of neural network model often used in image processing.
Cross Validation	A method of evaluating model performance by splitting the data into parts, taking one of these parts out to use as the test set, and averaging the results of the tests.
Data Annotation	See *Labelling*.
Deep Learning	A sub-type of Machine Learning, in which algorithms are built from layers of input-output units, allowing models to learn very complex patterns.
DistilBERT	A distillation of BERT which is 40% smaller but retains 97% of its performance (Sanh, Debut, Chaumond, & Wolf, 2019).
Domestic Abuse	In this thesis refers to Women's Aid definition: "We define domestic abuse as an incident or pattern of incidents of controlling, coercive, threatening, degrading and violent behaviour, including sexual violence, in the majority of cases by a partner or ex-partner, but also by a family member or carer." (Women's Aid, 2021b)

Embedding	E.g. GLoVE. A feature engineering technique in which words are encoded as a vector capturing their position in a multi-dimensional space, which represents the typical use of words in a language, and has been learnt from a very large corpus of text.
Encoding	Since computers are incapable of understanding natural language in its raw form, it is necessary to create a numeric representation of the text that can be used as input into the model. This is known as encoding.
Epoch	The number of epochs indicates the number of times that the model runs through all of the training data whilst training. More epochs give the model more exposure to the data, which may lead to higher accuracy but increase the risk of overfitting.
Exploratory Analysis	The process of computationally exploring and visualising datasets to understand their structure and content, before proceeding to more in-depth modelling.
F1 Score	Evaluation metric - weighted average between Precision and Recall (Alpaydin, 2020).
Feature Engineering	The way in which data is encoded and transformed before being given to a model as input has a significant impact on the way that the model learns and what its outcomes are. Feature engineering is concerned with optimising this process by extracting appropriate features from the data that will help the model to learn better (Alpaydin, 2020).

Few-shot Learning	The process of supervised machine learning models learning from only a small number of training examples (Wang, Yao, Kwok, & Ni, 2020).
Fine Tuning	The process of tweaking a pre-trained model to perform a specific task.
Hugging Face	A machine learning library that carries implementations of large machine learning models such as BERT and DistilBERT (Wolf et al., 2019).
Hyper-parameters	Macro parameters of a machine learning model that can be changed to adjust the way the model learns – for example, number of epochs, learning rate.
Imbalanced Dataset	Opposite of *Balanced Dataset*.
Instance	The term *instance* is sometimes used to refer to a single item of data in a dataset.
K-Means Clustering	An unsupervised algorithm which aims to split inputs into k clusters, by minimizing the distances between points in a cluster (Bhattacharya, Eube, Röglin, & Schmidt, 2019).
Labelling	A process in which a subset of available data is manually annotated with the correct output which the model should attempt to emulate for each data instance.
Large Language Models (LLMs)	Huge pre-trained models like OpenAI's GPT-4 which have learnt an understanding of natural language from enormous amounts of data scraped from the internet.
Latent Dirichlet Allocation (LDA)	A type of unsupervised topic modelling algorithm which tries to find related topics within unstructured text data.

Linear Support Vector Classification (SVC)	A type of machine learning algorithm which performs linear classification (see also *Support Vector Machine*).
Logistic Regression	A simple regression algorithm used in machine learning. Parameters are estimated by maximum-likelihood estimation (MLE).
Long Short Term Memory (LSTM)	A type of recurrent neural network often used in natural language processing tasks, which performs well at capturing relationships within sequences or series (e.g. sentences).
Machine Learning	A family of techniques broadly concerned with using complex algorithms to find patterns in large amounts of data.
N-gram	A combination of n words.
Overfitting	When a model fits to the training data very well but doesn't generalise to unseen data, this is known as overfitting.
Part-of-Speech tagging	When words in a sentence are automatically categorised according to their grammatical function (e.g. verb, subject, object etc.)
Pipeline	In machine learning, a pipeline refers to the computational set up of the steps surrounding the machine learning model – e.g. pre-processing, encoding, training, evaluation etc.
Precision	Specificity, or true positive rate.
Pre-processing	Steps taken to process text before feeding it into a machine learning model (e.g. removing common words, reducing words to their root form etc.)

Pre-trained Model	A machine learning model that has already been trained (learnt its parameters) on a large corpus of text and therefore can more easily be fine-tuned on tasks with only a small amount of training data.
Psychological Abuse	Despite the lack of a widely agreed-upon definition, there are nonetheless certain behaviours which are recognised as being characteristic of psychological abuse by DA advocates, researchers, psychologists and legal professionals, such as: consistently isolating a victim from their friends or family; undermining their sense of self through constant criticism; manipulating their sense of reality through denial or lying; not allowing them access to resources like money or technology; threatening them or their loved ones with physical violence; and many others (Home Office, 2022; SafeLives, 2019; Stark, 2009; Thompson, Basile, Hertz, & Sitterle, 2006). SafeLives, a DA charity in the UK, identifies that "perpetrators [of psychological violence] employ a wide range of psychological tactics, often personalised to the victim, to maintain control" (SafeLives, 2019) [Pg. 13].
Python	A widely used programming language popular in data science.
Random Forest	A type of machine learning algorithm based on decision trees.
Recall	Sensitivity or true negative rate.
Recurrent Neural Network (RNN)	A type of neural network which uses recurrent units that help to encode information held in sequences or series of information (e.g. sentences) (Alpaydin, 2020)
Scikit-learn	A widely used Python library for machine learning.

SpaCy	A commonly used Python coding package specifically designed for data science and machine learning applications (Spacy.io)
Supervised Learning	A type of Machine Learning which involves learning from datasets which have been labelled, meaning the inputs in the dataset that the model learns from (the training set) already have output labels assigned to them - for example, ‘contains abusive text’ or ‘does not contain abusive text’ are examples of labels. The labelling of the initial training set is often done by one or more human annotators. The algorithm can then use the patterns it learns from this dataset to appropriately assign or predict values for unseen or out-of-sample data which doesn’t have any existing labels.
Support Vector Machine (SVM)	A type of machine learning algorithm which performs linear classification (as well as non-linear classification using kernels) by maximising the margins between classes in a multi-dimensional space.
Test Set	A section of the data which is held out for testing, so that the model is tested on data that it hasn’t yet seen.
TF-IDF Vectors	A common method for encoding text that captures which words in a text are most relevant (Ramos, 2003). This is achieved by comparing the frequency of a word (or n-gram) in a particular text to its frequency in the whole dataset.
Tokenization	The splitting of a text into separate *tokens* (usually words) so that the text can be fed as individual pieces to a machine learning model.
Topic Modelling	An unsupervised learning approach that identifies related topics within a dataset of text.

Training data (or Training Set)	The section of the data which is fed to the machine learning model for it to learn from. (See also *Test Set*)
Train-test split	The split between training and testing data in the dataset (e.g. an 80-20 train-test split means that 80% of the labelled data is used for training and 20% is held back for testing).
Transfer Learning	The ability of models to learn characteristics of language from one dataset and apply these on a new dataset.
Transformer based models	A modern type of deep learning architecture which often underpins Large Language Models, because it requires less training time than RNN and CNN type models. Transformer models are based on an attention mechanism which allows the model to identify and prioritise the most relevant information in a contextual sequence (Vaswani et al., 2017).
Unsupervised Learning	Unsupervised learning involves learning from datasets without any ground-truth labels, where the learning comes not from existing labelled outputs but from the inherent structures within the data.
Validation set	In certain deep learning architectures, the train-test split is modified to include a third set, the validation set, which is used at each training step of the model to validate the results as training is occurring. This helps the model know how to update its parameters to move towards an optimal fit. The test set is then reserved to test the model at the end of training (Alpaydin, 2020).
Vector	A list of numbers, which can represent a position, or a quantity and direction, within a multi-dimensional space. Vectors are used in machine learning to represent and manipulate information.

Chapter 1: Introduction

Domestic Abuse (DA) is a broad term which encompasses a wide range of harmful and abusive behaviours, most usually perpetrated by a current or intimate partner, but sometimes by a family member or carer (Women's Aid, 2021b). DA includes, but is not limited to, physical, sexual, economic, psychological and emotional abuse (Women's Aid, 2021b). DA is an extremely widespread problem: The World Health Organisation estimates that 27% of women aged 15-49 years who have been in a relationship have experienced some form of intimate partner abuse (World Health Organisation, 2021). Data from the Crime Survey for England and Wales suggests that approximately 7.3% of women and 3.6% of men in the UK experienced DA in the year ending March 2020 (ONS, 2020).

Much of the existing large-scale data about DA is drawn from traditional survey- and questionnaire-based research (Australian Bureau of Statistics, 2013; Basile et al., 2011; European Union Agency for Fundamental Rights, 2014). Whilst such surveys are useful to understand DA on a population level, they are also costly, infrequent, and unlikely to capture granular data (Australian Bureau of Statistics, 2013). In this context, researchers often turn to interview-based approaches (Houston-Kolnik & Vasquez, 2022; Vatnar & Bjørkly, 2008). Although valuable, one-on-one interviews may also suffer from selection-bias, sample size issues, and being time-consuming to run (Karystianis et al., 2022).

Against this backdrop, some DA researchers are turning to secondary analysis of existing data (Australian Bureau of Statistics, 2013). Organisations that interact with victim-survivors - such as police forces or health services – collect large quantities of DA data which they are unable to analyse manually (Botelle et al., 2022; Karystianis et al., 2022). Additionally, victim-survivors of DA increasingly make use of online venues such as blogs and bulletin boards to express their experiences of abuse and to receive and offer support (Chu, Su, Kong, Shi, & Wang, 2021; S. Subramani et al., 2019). These entries generate huge amounts of text data, much of which is publicly accessible. Researchers and others working with victims of DA may want to leverage this text data to understand the experiences of victim-survivors.

Computational text mining is a set of techniques which use algorithms to understand, categorise or extract information from unstructured text data (DiMaggio, 2015). These can range from *simple* (for example, counting the occurrences of a pair of words in a corpus (Homan, Schrading, Ptucha, Cerulli, & Alm, 2020)) to *complex* approaches (for example, Deep Learning classifiers which use many layered neural networks to automatically categorise texts (S. Subramani et al., 2019)). Computational text mining methodologies have been used to harness big data to research social phenomena in other domains, such as the study of online hate (Fortuna & Nunes, 2018), cyberbullying (Hugo Rosa et al., 2019; Salawu, He, & Lumsden, 2020), right wing terrorism (Torregrosa, Bello-Orgaz, Martinez-Camara, Del Ser, & Camacho, 2021), and child abuse victimisation (Shahi et al., 2021). This thesis therefore seeks to understand whether, and if so, how, similar methods have been used to investigate and understand DA, and to develop a tool for detecting and understanding psychological abuse in a large dataset of online narratives about abuse.

1.1. Research Questions

This leads to the research questions of this thesis:

RQ1: How has existing work has used computational text analysis methods to research domestic abuse?

RQ2: How can we build on existing research to create a dataset of reported psychological abuse in online forums?

RQ3: How successfully do machine learning models learn to classify psychological abuse?

1.2. Thesis Outline

The remainder of the thesis seeks to answer these research questions through the following structure: Chapter 2 describes a systematic literature review, which identifies existing work using computational text analysis to research domestic abuse. This includes an assessment of the quality of the existing work, what kinds of data are used, and an identification of remaining gaps in the literature. Chapter 3 illustrates the process of creating a dataset of social media posts describing narratives of domestic abuse, annotated according to different types of psychological abuse mentioned in the posts. To create this dataset, a review of existing research and conceptualisations of psychological abuse was conducted and used as the foundation for a typology of psychological abuse which was refined through expert dicussion. Furthermore, the dataset was constructed according to *data explainability* principles in order to ensure its transparency and ethical use in downstream applications. Chapter 4 explores the use of computational text analysis methods on the dataset described in Chapter 3, including the training of machine learning models to automatically recognise different types of psychological abuse. This chapter presents the results of the classifiers and discusses future work which could be done to increase their performance. Finally, the Conclusion offers suggestions for future work and concluding remarks.

1.3. Contribution Statement

Sections of this thesis have been published, or are being prepared for publication, with multiple contributing authors, some of whom are current PhD students. This section clarifies the contribution of the thesis author, and where published works appear in the thesis.

Chapter 3, the systematic literature review, is adapted from the paper "A Systematic Literature Review of the Use of Computational Text Analysis Methods in Intimate Partner Violence Research" which was published in the Journal of Family Violence in March 2023. The thesis author was the first author of the paper and was responsible for: devising the review protocol; screening abstracts and full text of studies included in the review; and drafting, reviewing and editing the text. Isabel Straw (PhD Student) assisted with the screening of abstracts and full texts of studies included in the review. Dr. Mariconti and Dr. Tanczer provided supervision, reviewed the manuscript and provided minor text edits to the manuscript in their capacity as supervisors. The text in Chapter 3 of this thesis is copied directly from this paper with minor style edits to make it fit within the thesis; furthermore, Chapters 1 and 2 contain some sections of text that originally appeared in the background and introduction section of this paper.

Chapter 4, which describes the process of creating a labelled dataset of reports of psychological abuse, is currently being adapted for publication (outlet TBC). The expert discussions mentioned in Chapter 4, as well as the labelled dataset, were produced in collaboration with three other annotators: Lifang Li, Demelza Luna Reaver and Megan Knittel, the latter two being current PhD students. The annotators contributed to the discussions and labelling of the dataset, but did not participate in writing the text, all of which was drafted, reviewed and edited by the thesis author. Furthermore, the thesis author designed the study protocol, led the expert discussions, compiled the dataset, conducted the analysis, and conducted the literature review of existing measures of psychological abuse. Comments were received on the text of this Chapter from Dr. Mariconti and Dr. Tanczer as well as Dr. Mark Warner. Dr. Tanczer and Niamh Healy (PhD student) also contributed opinions to some of the expert discussions.

The thesis author was fully responsible conducting the remaining research and analysis, and for drafting, reviewing and editing the remainder of the text in this thesis. Comments and supervision were provided by Dr. Mariconti and Dr. Tanczer.

Chapter 2: Background

The following chapter provides an overview of existing relevant literature about domestic abuse (DA), particularly psychological abuse, and computational text analysis methods including machine learning. This provides background to the following analytical chapters.

2.1 Domestic Abuse

In popular culture, domestic abuse, often called domestic violence, has traditionally been perceived as referring to physical violence perpetrated by a current or former intimate partners (Women's Aid, 2023). However, it is now widely understood that non-physical abuse is more common in intimate relationships: Data from the Crime Survey for England and Wales (CESW) suggests that non-physical abuse is the most common type of DA - roughly 3% of the UK population had experienced non-physical abuse by an intimate partner in the past year during the period 2012-2020 (Home Office, 2021). Non-physical types of abuse include economic, sexual, technology-facilitated, psychological and emotional abuse (Women's Aid, 2021b) and these types often overlap.

Unfortunately, DA in intimate partner relationships is very common. Data from the World Health Organisation indicates that 27% of women worldwide aged 15-49 years who have been in a relationship have experienced some form of physical or sexual violence from an intimate partner (World Health Organisation, 2021). In the UK, national crime surveys in 2017 indicated that 14.8% of adults (of all genders) had experienced some form of non-sexual partner abuse since the age of 16 (Office for National Statistics, 2017). In England and Wales alone, two women a week are killed by a current or former partner (Office for National Statistics, 2018b). People of all genders and sexual orientations can be victims of DA, however, substantial evidence demonstrates that women experience DA much more often than men, and that it is most often perpetrated by men (Women's Aid, 2021b; World Health Organisation, 2023).

In addition, whilst DA most often occurs between current or former intimate partners, the term domestic abuse is used by some organisations and researchers to include abuse perpetrated by family members (sometimes referred to as Family Violence) or carers (Women's Aid, 2023). In this context, when practitioners want to be specific about abuse happening between current or former intimate partners, they sometimes use the term Intimate Partner Violence (IPV) or Intimate Partner Abuse (IPA) (World Health Organisation, 2023). This thesis uses the term Domestic Abuse (DA) according to the definition put forward by Women's Aid, a DA charity in the UK:

> *"We define domestic abuse as an incident or pattern of incidents of controlling, coercive, threatening, degrading and violent behaviour, including sexual violence, in the majority of cases by a partner or ex-partner, but also by a family member or carer."*
>
> (Women's Aid, 2021b)

This thesis is mostly focused on abuse between intimate partners, but the research contained within has also captured some abuse between family members, which is why the term Domestic Abuse (DA) is used throughout.

2.1.1. Definitional Difficulties

It is quite challenging to accurately measure the prevalence of DA, partly due to a non-homogenous set of definitions for what constitutes abuse across a wide variety of organisations that work with perpetrators and victim-survivors (B. Barocas, Emery, & Mills, 2016), particularly when it comes to non-physical forms of abuse (Dokkedahl et al., 2019). In addition, it is suspected that a large portion of abuse, particularly non-physical abuse, goes unreported due to shame, bias, unawareness of what constitutes abuse, or a victim-survivor's lack of access to services (Stark, 2009). In an EU-wide survey of 42,000 women conducted in 2012, 66% of respondents who had experienced DA did not report it to, or seek help from, any organisation, governmental or otherwise (European Union Agency for Fundamental Rights, 2014). Similarly, in the UK in March 2018, 82% of women who had experienced DA in the last year had not reported it to the police (Women's Aid, 2021a). For these reasons, understanding the true prevalence and presentations of different kinds of abuse remains an active research question within the DA research domain.

2.1.2. Psychological Abuse

Psychological abuse is a particularly widespread form of abuse (European Union Agency for Fundamental Rights, 2014), and has been recognised and studied as a distinct subtype of DA since at least the 1990s (Follingstad, Coyne, & Gambone, 2005). A US-wide survey in 2010 found that nearly half of all 16,507 male and female respondents had experienced at least one form of 'psychological aggression' by an intimate partner during their lifetime (Basile et al., 2011). Many physically violent relationships are also psychologically abusive (Dobash, Dobash, Wilson, & Daly, 1992; Johnson & Leone, 2005), and psychological abuse can have equally serious negative effects on mental health as physical abuse (Lagdon, Armour, & Stringer, 2014; Lawrence, Yoon, Langer, & Ro, 2009; Pico-Alfonso et al., 2006).

However, psychological abuse is still not as well understood as physical abuse, partly because it is more difficult to define (Dokkedahl et al., 2019; Follingstad, 2009). There are still ongoing debates around what constitutes psychological abuse, how to measure its severity, and even what to call it (Dokkedahl et al., 2019). For these reasons, many research questions about psychological abuse still persist, such as how common it is, how different groups experience it, and what interventions could be most helpful to victim-survivors of this kind of abuse (Dokkedahl et al., 2019; Follingstad, 2009; Lagdon et al., 2014). As Lagdon et al. describe, a "lack of clear validated measures assessing the impact of psychological violence has meant that researchers haven't clearly focused on this type of violence" [Pg. 7] (Lagdon et al., 2014).

Whilst it is difficult to accurately measure psychological abuse due to varying definitions, there is little doubt that it is likely to be widespread and extremely harmful to victim-survivors. Experience of psychological aggression from an intimate partner is associated with symptoms of depression and anxiety (Lawrence et al., 2009) and low self-esteem (Sackett & Saunders, 1999). Some studies have found that experience of psychological abuse is a *stronger* predictor of Post-Traumatic Stress Disorder (PTSD) in victim-survivors of intimate partner violence than experience of physical or sexual abuse (Norwood & Murphy, 2012; Taft, Murphy, King, Dedeyn, & Musser, 2005). Experiencing psychological abuse can certainly be as damaging to victim-survivor's mental health as experiencing physical abuse, although separate causalities can be difficult to untangle since physical and psychological abuse often co-occur (Lagdon et al., 2014; Lawrence et al., 2009; Pico-Alfonso et al., 2006).

Despite the lack of a widely agreed-upon definition, there are nonetheless certain behaviours which are recognised as being characteristic of psychological abuse by DA advocates, researchers, psychologists and legal professionals, such as: consistently isolating a victim from their friends or family; undermining their sense of self through constant criticism; manipulating their sense of reality through denial or lying; not allowing them access to resources like money or technology; threatening them or their loved ones with physical violence; and many others (Home Office, 2022; SafeLives, 2019; Stark, 2009; Thompson et al., 2006).

SafeLives, a DA charity in the UK, identifies that "perpetrators [of psychological violence] employ a wide range of psychological tactics, often personalised to the victim, to maintain control" (SafeLives, 2019) [Pg. 13]. The contextual nature of psychological abuse makes it particularly challenging to identify (Crown Prosecution Service, 2017). Surveys or questionnaires which aim to measure psychological abuse tend to include some variation of the types of behaviours mentioned above, but may fail to capture more nuanced or personalised forms of abuse (Basile et al., 2011; European Union Agency for Fundamental Rights, 2014). Understanding psychological abuse from the perspective of the victim-survivor is also important, since psychological abuse is tailored to the victim's psycho-emotional context and may be perceived differently by outside observers (Stark, 2009). Survivor narratives are therefore particularly important when studying psychological abuse.

2.1.3. Using Surveys to Measure Psychological Abuse

Existing measures for researching psychological abuse at a population level have mostly involved population-based surveys. Comparing three large-scale surveys about abuse conducted in the US, EU and UK (Basile et al., 2011, European Union Agency for Fundamental Rights, 2014, Home Office, 2021) begins to offer insight into some of the many methodological issues around measuring and defining psychological abuse.

Each survey conceptualised and captured psychological abuse differently. All three surveys included questions about isolation from family and friends, humiliating or belittling treatment, economic control, and monitoring whereabouts (Basile et al., 2011; European Union Agency for Fundamental Rights, 2014; Home Office, 2021). The Crime Survey for

England and Wales (CSEW) also included a question about monitoring communications which was not included in the other two surveys. Both the US and the EU survey also included questions about threats of physical harm, threats to children, not allowing their partner to leave the house, and intimidating angry behaviour, which were not included in the CSEW. Finally, the EU survey included questions about suspicious and jealous behaviour and public humiliation which were not included in the US survey or the CSEW, and the US survey included questions about suicide threats, threats involving pets, destroying personal belongings, name calling and specific statements like: "told partner no one else would want them", "told partner they were a loser, a failure or not good enough" and "said things like 'if I can't have you then no one can'".

As can be seen from this comparison, each survey asked about a different set of psychologically abusive behaviours. As a result, the surveys are likely to capture slightly different but overlapping phenomena, which makes them difficult to compare. In addition, overall prevalence statistics in each study were based on respondents answering “yes” to having experienced any one or more of these behaviours in an intimate partner relationship. However, as emphasised by Follingstad (Follingstad, 2007) when discussing the definition of ‘psychological abuse’, it is important to distinguish between a single ‘abusive’ behaviour and between a pattern of these behaviours that crosses the moral threshold for ‘abuse’. Otherwise, there is a risk of criminalising unpleasant but extremely common acts of psychological aggression which often occur, as part of the rich emotional tapestry of intimacy, in otherwise happy and healthy partnerships (Follingstad, 2007).

2.1.3. Existing Measures of Psychological Abuse

These methodological difficulties in measuring psychological abuse extend beyond large-scale population studies. Recent work from Dokkedahl et al. (Dokkedahl et al., 2019) systematically reviewed existing psychometric measures of psychological abuse used in DA research and practice. This resulted in a collection of 20 psychometric measures of psychological violence (Dokkedahl et al., 2019), themselves collated by combining a compendium from the Center for Disease Control and Prevention (Thompson et al., 2006) with research from the National Unit against IPV in Denmark, LUV (Oldrup, Andersen, Kjær, Nielsen, & von Rosen, 2018). The variety of the 20 measures, and the diversity of behaviours included in their respective questionnaires, underlines the difficulty in pinning down a conclusive definition of psychological abuse. However, examining this collection of measures provides a proxy to understand, on average, how researchers and practitioners have conceptualised and measured psychological abuse over the past decades.

In essence, there is no widely agreed upon definition, measure or threshold for what constitutes psychological abuse, which makes it a difficult phenomenon to study. There is therefore a need for further research into how people experience and report psychologically abusive behaviours. Categorical surveys investigating psychological abuse are in some way limited, since the questions they ask inherently encode a certain understanding of psychological abuse and may inadvertently exclude relevant behaviours.

2.2. Computational Text Analysis and DA Research

In this context, social media platforms can provide rich sources of data for studying social and criminal phenomena (Baden, Pipal, Schoonvelde, & van der Velden, 2022; DiMaggio, 2015). Studying human behaviour in organic, online social spaces can act as a complementary method to traditional survey- and questionnaire- based research (Alvarez, 2016; Lazer et al., 2009). It allows for the study of spontaneous expressions of experience, capturing data beyond potentially narrow survey questions, and avoids potential participant social desirability bias (where a participant is influenced to answer a survey in a certain way because they know their response is being recorded by a researcher) (Alvarez, 2016). In this way, using social media data to study reports of psychological abuse could provide a useful augmentation to traditional survey- or questionnaire- based research.

Computational text analysis methods, which use computer algorithms to understand large amounts of text data, have been successfully used to study a wide range of topics and harms including online hate speech (Fortuna & Nunes, 2018; Schmidt & Wiegand, 2017; Waqas, Salminen, Jung, Almerekhi, & Jansen, 2019), cyberbullying (Chatzakou et al., 2017; Kim, Razi, Stringhini, Wisniewski, & De Choudhury, 2021; H. Rosa, Matos, Ribeiro, Coheur, & Carvalho, 2018; Hugo Rosa et al., 2019; Salawu et al., 2020), right wing terrorism (Hartung, Klinger, Schmidtke, & Vogel, 2017; Torregrosa et al., 2021), and child abuse victimisation (Amrit, Paauw, Aly, & Lavric, 2017; Annapragada, Donaruma-Kwoh, Annapragada, & Starosolski, 2021; Babvey et al., 2021; Shahi et al., 2021). The success of computational methods for research in these domains suggest that they may be useful for the study of DA.

Chapter 3 will survey existing research using computational text analysis methods in DA research, to discover how such methods are already being used and identify gaps in the literature. However, to provide background for this interdisciplinary thesis, it is first necessary to survey literature about computational text analysis methods, including machine learning, to identify and explain relevant techniques and tools that will be mentioned in future chapters.

2.3. Machine Learning

Computational text analysis is an umbrella term which includes a wide range of computational methods and algorithms used to study large datasets of text. Some of these methods involve relatively simple exploratory analysis, the process of computationally exploring and visualising datasets to understand their structure and content, before proceeding to more in-depth modelling. For example, Homan et al. explored the most frequent n-gram word collocations (unigrams, trigrams and bigrams) in a dataset of Tweets using the #WhyIStayed and #WhyILeft hashtags (Homan et al., 2020). Homan et al. also used another exploratory technique, Part-of-Speech tagging, where words in a sentence are automatically categorised according to their grammatical function, to identify the most common Subject-Verb-Object combinations in their dataset.

Other methods used in computational text analysis are examples of machine learning (ML), which is a family of techniques broadly concerned with using complex algorithms to find patterns in large amounts of data. ML has applications in a wide range of tasks, but generally speaking these tasks tend to be classified into either regression or classification tasks, where regression tasks are concerned with predicting or assigning a continuous output value for an input, whilst classification tasks are concerned with categorising and input into one or more output classes (Prabakaran, Waylan, & Penfold, 2017).

2.3.1. Types of Machine Learning

Traditional ML regression algorithms include Logistic Regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression (Tibshirani, 1996), whereas classification algorithms include Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Decision Tree (DT) (Alpaydin, 2020). It is beyond the scope of this thesis to explain the mechanisms behind these algorithms, but clear introductory explanations can be found in Prabakaran et al. (Prabakaran et al., 2017).

Deep learning is a sub-type of ML, in which algorithms include layers of input-output units which allow the model to learn more complex patterns (Alpaydin, 2020). Supervised deep learning models include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), as well as Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014) models, both specific types of RNNs that perform well in natural language tasks. It is important to note that Deep Learning models tend to be highly complex and opaque, meaning that it is difficult for humans to understand their decision-making mechanisms. When working in sensitive or controversial areas, this may lead to a problematic lack of transparency (Samek, Wiegand, & Müller, 2017; Xie, Ras, van Gerven, & Doran, 2020).

2.3.2. Supervised Learning

ML techniques can be further categorised into supervised and unsupervised learning. Supervised learning involves learning from datasets which have been labelled, meaning the inputs in the dataset that the model learns from (the training set) already have output labels assigned to them - for example, ‘contains abusive text’ or ‘does not contain abusive text’ are examples of labels. The labelling of the initial training set is often done by one or more human annotators. The algorithm can then use the patterns it learns from this dataset to appropriately assign or predict values for unseen or out-of-sample data which doesn’t have any existing labels.

Supervised learning has already been used in some DA research. For example, Victor et al. train a classifier on a corpus of child welfare investigation summaries, which were manually labelled as to whether or not the child was in need of domestic violence service intervention (Victor, Perron, Sokol, Fedina, & Ryan, 2021). Their supervised classification model achieved high accuracy, which led them to conclude that "insights derived from these procedures can be particularly useful for investigating the prevalence, temporal trends and geographic distribution of domestic violence-related needs"[Pg. 1] (Victor et al., 2021).

Whilst supervised learning is an extremely powerful tool, it comes with potential pitfalls. Supervised classification problems require a training dataset with particular properties to work well. Firstly, it is desirable that the dataset be balanced, which means that it contains relatively equal numbers of instances from each 'class' (category). An unbalanced dataset makes it harder for the model to discriminate between classes, since it tends to learn more from the characteristics of the dominant class (Akbani, Kwek, & Japkowicz, 2004). Secondly, to achieve good results with classification, a dataset ideally contains a clear separation between classes. This may not be a linear separation, since many models operate in very high dimensional spaces, but some classes may simply have too many overlapping properties which can make them very difficult to separate (Alpaydin, 2020).

Finally, the manual labelling method of the training dataset is of paramount importance in supervised classification, since it directly informs the input-output relationship that the model will learn. Any bias or inaccuracies in the labelling process are likely to be picked up and replicated by the model (Dignum, 2017, Bechmann and Zevenbergen, 2019).

2.3.3. Data Annotation For Machine Learning

The annotation of datasets to provide a ground truth for training supervised learning models presents a number of potential pitfalls. Firstly, data annotation is time consuming and laborious work (Muller et al., 2021). Secondly, it may be difficult for human annotators to agree on exactly what is the ground truth label for a piece of data, especially when the subject matter is nuanced or contextual (Kulesza, Amershi, Caruana, Fisher, & Charles, 2014). When multiple annotators engage in debate or discussion about how to apply labels, this process is often hidden from view in the final dataset, making the assumptions made when labelling the data opaque to downstream users (Muller et al., 2021). An emerging body of work addresses best practices for annotating machine learning datasets to avoid or ameliorate some of these issues (Kapania, Taylor, & Wang, 2023; Prabhakaran, Davani, & Diaz, 2021; Röttger, Vidgen, Hovy, & Pierrehumbert, 2021), which will be investigated further in Chapter 4.

2.3.4. Pre-trained Models

Due to the many difficulties encountered in manually labelling data, supervised learning datasets are often relatively small, which can limit the performance of the models which train on them. However, in recent years, the emergence of large pre-trained language models has significantly improved the state-of-the-art performance in 'few-shot learning', meaning the ability of supervised learning models to learn from only a small number of training examples (Sun, Qiu, Xu, & Huang, 2019). This has been achieved through the use of 'transfer learning', which is the ability of models to learn characteristics of language from one dataset and apply these on a new dataset (Ge, Guo, Das, Al-Garadi, & Sarker, 2023).

Large pre-trained models, such as BERT (Devlin, Chang, Lee, & Toutanova, 2018), are trained on very large corpuses of data taken from the internet. This allows them to learn a general representation of the English language. Such models can then be ‘fine-tuned’ on custom classification tasks, and since they already have some ‘understanding’ of language, it is easier for them to learn from a small number of examples (Sun et al., 2019).

Since BERT is a very large model, training it can be prohibitive in terms of computing resources. Smaller versions of these large models have been created which allow users with fewer computing resources to train the model with only minimal reductions in accuracy. DistilBERT is a distillation of BERT which is 40% small but retains 97% of its performance (Sanh et al., 2019).

2.3.5. Unsupervised Learning

In contrast to the supervised learning approaches discussed so far, unsupervised learning involves learning from datasets without any ground-truth labels, where the learning comes not from existing labelled outputs but from the inherent structures within the data (Alpaydin, 2020) [Pg. 11]. A common example of unsupervised learning is clustering algorithms, which identify related clusters in data. Using unsupervised learning for text can, for example, help to identify topics within large amount of text data – this approach is known as topic modelling (Nikolenko, Koltcov, & Koltsova, 2017). For example, Xue et al. used a popular topic modelling technique, Latent Dirichlet Allocation (LDA), to extract 9 themes from over 1 million tweets about family violence during the COVID-19 pandemic (Xue, Chen, Chen, Hu, & Zhu, 2020). Chu et al. also LDA clustering to extract the most common topics and their associated words from different types of support groups in a Chinese forum offering support to IPA victim-survivors, finding themes of ‘emotional support’ and ‘informational support’ (Chu et al., 2021).

2.3.5.1. K-means clustering

One method for unsupervised topic modelling is k-means clustering. K-means clustering aims to split inputs into k clusters, by minimizing the distances between points in a cluster (Lloyd, 1982). This is a difficult problem to solve computationally (it is known as an “NP-Hard” problem in computer science) (Aloise, Deshpande, Hansen, & Popat, 2009), but numerous algorithms have been proposed that allow fast convergence to a local optimum (Arthur & Vassilvitskii, 2007), meaning they find a good solution quickly, but don’t guarantee finding the best possible existing solution. K-means clustering is a useful method for experimentally exploring a dataset and seeing if clusters emerge, although for a sparse and complex text dataset clusters may not fully converge, or may be difficult to interpret (Nikolenko et al., 2017).

2.3.6. Feature Engineering and Pre-Processing

All the ML models mentioned so far need to be trained on data, but the way in which the data is encoded and transformed before being given to a model as input has a significant impact on the way that the model learns and what its outcomes are (Heaton, 2016). The processing of data input into a format that can be fed into an ML model has three inter-related stages: pre-processing, encoding and feature engineering.

In traditional text-based ML, pre-processing commonly involves steps such as data cleaning (e.g. removing URLS), removing stop-words (removing common words such as “a” and

“and” that appear in almost all data instances), stemming and/or lemmatisation (turning words into their grammatical stems or lemmas e.g. detecting -> detect).

Feature engineering is concerned with extracting from the data appropriate features that will help the model to learn more robustly – for example: counts of particular keywords within a text; sentiment analysis scores (how ‘positive’ or ‘negative’ a piece of text is according to a pre-built sentiment model); or part-of-speech tags (which identify which words are verbs, adjectives etc.). Adding this extra information to the encoding of text instances that are fed into a model can help it to learn patterns in the data more effectively. Especially when dealing with sparse data – small datasets or unbalanced datasets, that is, datasets where the number of instances in each class is not equal - feature engineering can be an extremely important way of improving performance (Duboue, 2020; Heaton, 2016). Furthermore, analysing different types of features and how they impact model behaviour can help to provide qualitative insight into the model and the problem at hand.

2.3.6.1. TF-IDF Vectors

Since computers are incapable of understanding natural language in its raw form, it is necessary to create a numeric representation of the text that can be used as input into the model. This is known as encoding. A common method for encoding text is using Term Frequency-Inverse Document Frequency (TF-IDF) vectors – for example, Victor et al. used TF-IDF vectors as input into their model which classified domestic violence in child welfare records (Victor et al., 2021). TF-IDF vectors are a way of capturing which words in a text are most relevant (Ramos, 2003). This is achieved by comparing the frequency of a word (or n-gram) in a particular text to its frequency in the whole dataset. Intuitively, this means that an uncommon word that appears frequently in only a few posts will have more importance than a common word that appear frequently in most posts (Ramos, 2003).

2.3.6.2. Word Embeddings

TF-IDF vectors can be directly fed into a ML model as input. However, to enrich the information contained within the inputs, a common feature engineering technique is the use of word embeddings. Word embeddings, such as GLoVE (Pennington, Socher, & Manning, 2014) and Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013), are extremely useful in NLP tasks because they encode information not only about the frequency of words in a text, but also about the words' relationship to each other within a language. The creation of a word embedding exploits the assumption that words with the same context are likely to have related meaning (Almeida & Xexéo, 2019). Words are encoded as a vector capturing their position in a multi-dimensional space, which represents the typical use of words in a language, and has been learnt from a very large corpus of text (Almeida & Xexéo, 2019; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2017). These vectors can then be used as inputs into ML models, and often result in better performance since the vectors themselves already capture a lot of information about language (Mikolov et al., 2017).

2.3.7. Evaluation

Once models have been trained, there needs to be a way of evaluating their performance – how well do they manage the intended task on unseen data that was not used during training? This question is vital to avoid overfitting models, where models learn the characteristics of their training set "too well" in a way that means their results don't generalise to other data (Arango, Pérez, & Poblete, 2019). Classification models can be tested on a test set, which is a portion of the dataset that is set aside during training as a test-bed to check the algorithm's performance on out-of-sample data for which the researcher already knows the correct labels. In cases of data sparsity, a mechanism called k-fold cross validation is often used to check model performance, in combination with or instead of a separate test set. This involves separating the labelled data into k-classes (or folds), in sequence taking one of these classes out to use as the test set, and averaging the results of these runs (Alpaydin, 2020).

When reviewing ML literature it is important to take into account not only the model architecture but also the dataset, the pre-processing and feature engineering methods, and the evaluation mechanism, since these can have as much impact on the model performance and its application as the model architecture itself.

2.4. Conclusion

This section has introduced psychological abuse as a sub-type of DA that is still not well defined or understood, and computational social science methods, including machine learning, as a potential contribution to research about psychological abuse in large text datasets.

Chapter 3: A Systematic Literature Review of the use of Computational Text Analysis Methods in Domestic Abuse Research

3.1. Introduction

As discussed in Chapter 2, existing methods for studying domestic abuse (DA) draw largely from the social sciences. These include primary data collection tools such as surveys (Lagdon et al., 2022; Mahoney, Farrell, & Murphy, 2022; ONS, 2020), interviews or focus groups (Avieli, 2021; Øverlien, Hellevik, & Korkmaz, 2020; Wood, Backes, Baumler, & McGiffert, 2021), as well as secondary analyses of data sourced from, for example, DA service providers (Rogers, Rumley, & Lovatt, 2019).

Furthermore, Chapter 2 introduced how recent developments in the field of computational social science have led to data science tools which extend and complement established social science techniques (DiMaggio, 2015; Evans & Aceves, 2016). These tools further ease the data collection and analysis process by harnessing big data and Machine Learning (ML) (Casquilho-Martins, Belchior-Rocha, & Moro, 2022; Gauthier & Wallace, 2022).

Indeed, a small number of studies have applied such computational approaches to the study of DA. Publications examined online support-seeking behaviours of victim-survivors (Chu et al., 2021), studied reasons given for staying and leaving abusive relationships in microblog posts (Schrading, Alm, Ptucha, & Homan, 2015b), and identified crisis posts on social media platforms such as Facebook (S. Subramani, Wang, Vu, & Li, 2018). In addition, computational methods have offered DA researchers access to datasets which are simply too large to evaluate manually e.g., police incident reports (J. Poelmans, Elzinga, Viaene, Hulle, & Dedene, 2009; Wilson, Spike, Karystianis, & Butler, 2021), case summaries (Victor et al., 2021), and Google search histories (Zaman et al., 2021).

Despite this small but growing body of work, there is yet no review addressing the application of computational text analysis methods to the study of DA. This omission stands in the way of proposing further methodological innovation, and to opening the field to the latest transdisciplinary research approaches stemming from computer science. This chapter seeks to fill this gap by conducting a systematic literature review of eight online academic databases (Scopus, ProQuest, Web of Science, IEEE Explore, PsychInfo, PubMed, ArXiv.org and ACM Digital Library).

The rest of the chapter is structured as follows: 1) Research Questions: A number of RQs are proposed to investigate the use of text mining methods in the DA domain. 2) Methodology: The methodology of this review is described, including the search strategy and inclusion criteria. 3) Results: The results of the review are summarised and analysed using a 21-item checklist 4) Discussion: The findings from the review, its limitations, and potential directions for future work are discussed. 5) Concluding remarks.

3.2. Research Questions

This chapter offers a systematic review of existing work which has applied computational text mining to the study of DA. In doing so, it aims to provide a resource for DA scholars who may want to use computational text methodologies in their work, providing a starting point to understand current capabilities as well as directions for future research. The chapter gives an introductory background to text mining methods and techniques, whilst seeking to examine the quality of current work. No existing knowledge of computational methodology is assumed, and all terminology will be explained within this chapter.

The assessment of the academic literature is driven by three research questions: (RQ1) How have computational text analysis methods been *used* in DA research?; (RQ2) What datasets are *available* for studying DA using computational text analysis?; (RQ3) How have text analysis methods been *evaluated* in the study of DA?

3.3. Method

A systematic review of existing academic literature was conducted according to PRISMA-P guidelines (Moher et al., 2015) between November 2021 and July 2022.

3.3.1. Electronic Search Strategy

Eight databases (ACM Digital Library, ArXiv.org, IEEE Xplore, ProQuest, PsychInfo, PubMed, Web of Science, Scopus) were searched for records containing *both* terms relating to computational text mining *and* terms relating to intimate partner violence, within all fields apart from the full-text (e.g. Title, abstract, keywords, publication venue)[1]. The full search string was as follows:

> *(("artificial intelligence" OR "machine learning" OR "supervised learning" OR "unsupervised learning" OR "automatic detection" OR "automatic recognition" OR "text mining" OR "natural language processing" OR "deep learning" OR "text analysis" OR "information retrieval" OR "information extraction" OR "machine reading" OR "word embeddings" OR "feature extraction" OR "knowledge discovery" OR "data engineering" OR "knowledge engineering" OR "exploratory data analysis" OR "quantitative content analysis" OR "automatic content analysis" OR "computational methods" OR "big data" OR "predictive model") AND ("intimate*

[1] NB The search string was adapted to fit the search functions of different databases. In ArXiv.org, only the DA-related part of the search string was used, since all research on ArXiv.org was assumed to have a computational element and the search function did not allow for so many search terms.

partner violence" OR "intimate partner abuse" OR "domestic violence" OR "domestic abuse" OR "family violence" OR "family abuse"))

3.3.2. Inclusion Criteria

Studies were included in the review if they met the following criteria:

- Peer reviewed and pre-print academic literature;
- The study uses computational text analysis or text mining to address an DA-related outcome from a large (n>50) dataset which includes unstructured text fields;
- The study includes results from at least one dataset (studies which discuss a purely theoretical design or prototype were excluded);
- The main outcome of the computational model is related to the identification of types, characteristics, prevalence, behaviours and/or opinions of DA (We excluded studies where DA is used as an input feature rather than an outcome, for example studies measuring the impact of DA (input) on mental health (outcome));
- Since DA is defined differently in different research, and sometimes is captured within other definitions of violence, studies with “family violence” “intimate partner violence” or “sexual violence” related outcomes were included, since these may include DA within their definitions.

3.3.3. Data Extraction and Management

Records identified through database searches were imported into Rayyan (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016) for data management. After duplicates had been discarded, two of the authors independently performed abstract screening according to the above inclusion criteria. Cohen’s Kappa statistic was calculated to determine Inter Rater Reliability (IRR) following the procedure described by Hallgren (Hallgren, 2012). Cohen’s

Kappa was 0.69, indicating a substantial level of agreement between the two reviewers according to guidelines from Landis and Koch (Landis & Koch, 1977). Remaining disagreements were resolved following a discussion between the two reviewers.

The included papers were subsequently downloaded and a pro-forma was used to extract the information from each paper. The pro-forma was piloted with 16 initial papers and feedback was obtained from other authors, following which amendments were made. The final pro-forma consisted of the following information fields:

> *Authors; Name of study; Year of study; DA-related hypothesis or outcome; Source, size and time period of dataset; Demographics of dataset (if discussed); Method and results of labelling dataset; Data pre-processing and cleaning process (if mentioned); Feature selection process (if mentioned); Model task; Types of models tested; Best performing model; Evaluation method; Evaluation metrics used; Best evaluation outcome; Summary of discussion of evaluation outcomes (if any); Summary of interpretability of the model (if discussed); Technologies mentioned; The definition of violence used by the study (if any); Summary of ethical discussion or limitations (if any); Whether any code/datasets are open source.*

3.3.4. Quality Assessment

Existing guidelines for assessing bias, quality, and reliability of biomedical or psychological studies are difficult to apply to research using computational text-analysis methods. Particularly when reviewing highly specialised systems, such as those involving ML. This paper builds on existing frameworks for assessing ML and mixed methods research (Dreisbach, Koleck, Bourne, & Bakken, 2019; Hinds, Parkhouse, & Hotchin, 2021; Hong et al., 2018; Siebert et al., 2020; Zhai, Yin, Pellegrino, Haudek, & Shi, 2020) to develop a checklist of 21 ‘yes/no’ criteria which were used to assess the overall quality of studies included in the review. A wide range of approaches are surveyed in the included studies, so some irrelevant items were excluded from the checklist depending on the study in question. For that reason, the checklist is not supposed to provide a ranking of studies but an indication of overall quality of the included works.

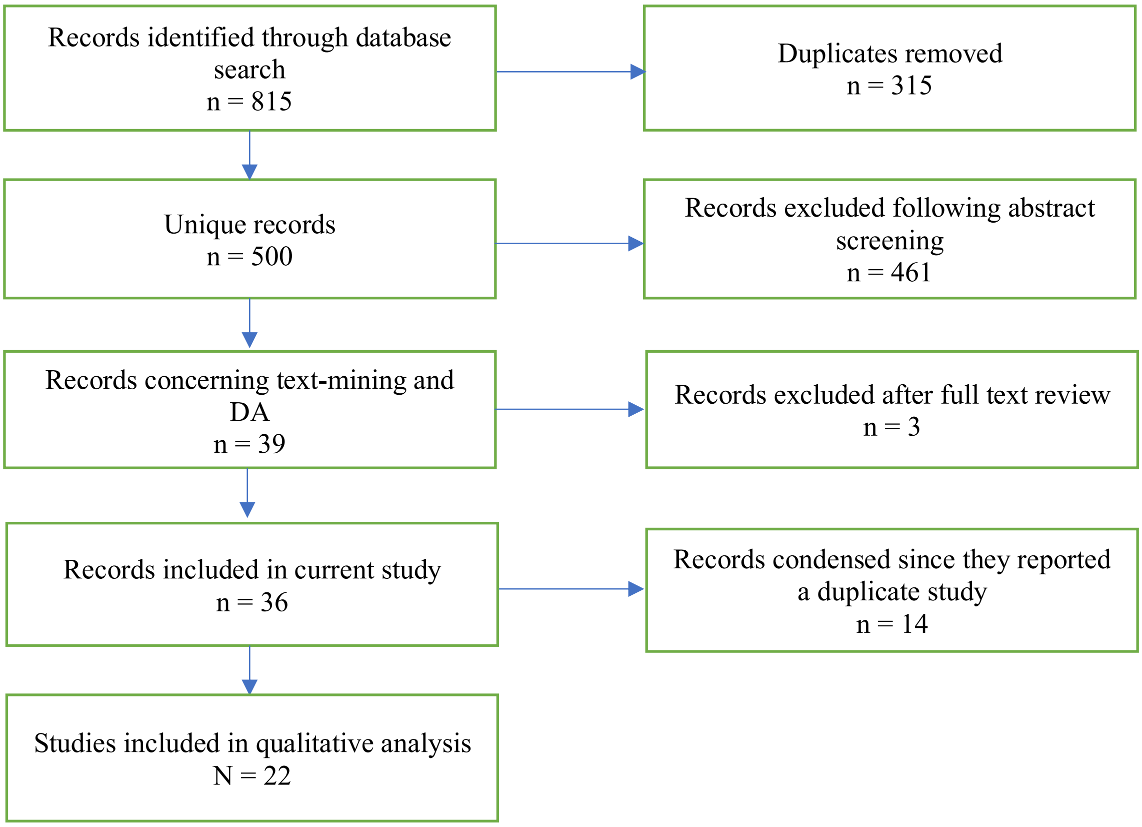
Checklist for Assessing the Quality of Included Studies:

1. Definition of violence discussed
2. Clearly described and motivated DA-related hypothesis or outcome
3. Representativeness/demographics of dataset discussed and/or analysed
4. Source, size and time period of dataset reported
5. Data cleaning and sampling process reported
6. Discussion of pre-processing techniques
7. Appropriate model used for hypothesis
8. Feature selection discussed and/or different features considered
9. Different models tested and compared
10. Clear and appropriate evaluation criteria
11. Evaluation outcomes reported
12. Evaluation outcomes discussed e.g. comparison to other work, discuss misclassifications
13. Study includes discussion of model interpretability, or clearly explains model rules
14. Includes ethical discussion
15. Source code and/or datasets available
16. Includes discussion of limitations of model and/or appropriate use
17. Dataset is of an appropriate size, and balance of classes discussed
18. Data labelling process is explained
19. Data is labelled according to a protocol by more than one annotator and IAA reported
20. Model is tested on held-out 'test' set
21. Model is tested or deployed "in the wild"

3.3.5. Included Studies

As can be seen in the PRISMA chart in Figure 3, the search yielded 815 results of which 315 were duplicates, leaving 500 unique studies. Of these, 461 were excluded as irrelevant (meaning they did not mention domestic abuse and/or use a computational text mining methodology) during abstract screening, leaving 39 papers.

Figure 3 - ***PRISMA Chart***


Records identified through database search
n = 815
Duplicates removed
n = 315
Unique records
n = 500
Records excluded following abstract screening
n = 461
Records concerning text-mining and DA
n = 39
Records excluded after full text review
n = 3
Records included in current study
n = 36
Records condensed since they reported a duplicate study
n = 14
Studies included in qualitative analysis
N = 22

Following full text review, a further three records were excluded because: no full text was available (n=1); the text was not written in English (n=1); the paper discussed a purely theoretical approach which did not involve any data (n=1). Finally, a number of papers (n=16) were found to report on the same two broad studies, using similar datasets and models. These were the Karystianis et al. papers on the New South Wales Police Force data using a rule-based approach, n=6 (Adily, Karystianis, & Butler, 2021; Hwang et al., 2020;

Karystianis et al., 2019; Karystianis et al., 2022; Wilson et al., 2021; Withall et al., 2022), and the Poelmans et al. papers on the Amsterdam-Amstelland Police Force Data using an FCA and ESOM based approach, n=10 (Elzinga, Poelmans, Viaene, & Dedene, 2009; J. Poelmans, Elzinga, & Dedene, 2013; J. Poelmans, Elzinga, Viaene, & Dedene, 2008, 2009; Jonas Poelmans, Elzinga, Viaene, & Dedene, 2010; J Poelmans, Elzinga, Viaene, & Dedene, 2011; J. Poelmans, Elzinga, Viaene, Dedene, & Van Hulle, 2009; J. Poelmans, Elzinga, Viaene, Hulle, et al., 2009; J. Poelmans, Elzinga, Viaene, Van Hulle, & Dedene, 2009; J Poelmans, Van Hulle, Viaene, Elzinga, & Dedene, 2011)). For simplicity of reporting in this review, these records were condensed into two unique studies. This left N=22 unique studies to be included in the following qualitative analysis. A summary of the included studies can be found in Table 1.

3.4. Results

The N=22 included studies cover a wide range of research questions and text mining methodologies. Outcomes include extracting topics from a corpus of social media texts (More & Francis, 2021; Rodriguez & Storer, 2020; Xue et al., 2020; Xue, Chen, & Gelles, 2019), information retrieval of abuse and injury types from police reports (Adily et al., 2021), detecting the presence or absence of mentions of domestic violence in various types of text (Botelle et al., 2022; Schrading et al., 2015b), and event and entity recognition from court documents (Li, Sheng, Ge, & Luo, 2019) and victim-survivor narratives (Y. Liu, Li, Liu, Zhang, & Si, 2019). A summary of the studies can be found in Table 1.

The quantity of this research seems to be increasing in recent years, with the majority (n=18) of studies being published in the last 5 years, and almost a third (n=7) being published in the last two years. This may be a reflection the increased public awareness of the ‘shadow pandemic' of domestic abuse brought on by the COVID-19 pandemic (Xue et al., 2020). Given the interdisciplinarity of the topic, it is interesting to note that there was an equal split between studies published in computer science journals and conferences[2] (n=11), and those published in social science and health related venues [3] (n=11).

The following section reviews the included studies as follows: firstly, by giving an overview of the different text mining models and techniques used in the studies; secondly, by reviewing the characteristics of the various datasets which studies used; and finally, by discussing how studies evaluated their techniques and models and what the evaluation outcomes were. This is followed by the Discussion section which investigates the quality of the included studies, offers lessons for researchers hoping to use text mining in their own work, considers ethical concerns of using computational text mining in the study of DA, and examines the limitations of the current review.

[2] e.g. NAACL, IEEE Transactions, Databases Theory and Applications

[3] e.g. Violence Against Women, Journal of Interpersonal Violence, Journal of Medical Internet Research

Table 1

Summary of Included Studies

Name of Study	Authors	Year	Dataset	Model Task	(Most Successful) Model Type	Evaluation Outcome (of best performing model)
Public Attention and Sentiment toward Intimate Partner Violence Based on Weibo in China: A Text Mining Approach	Xu et al.	2022	Chinese Social Media Comments about Yuya IPV disclosure	Unsupervised Sentiment Analysis	Custom rule-based sentiment analysis algorithm using a combination of pre-trained and custom-built dictionaries	N/A - unsupervised classification
Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study.	Botelle et al.	2022	Electronic Health Records (EHRs) from the South London and Maudsley NHS Foundation Trust	Seven binary classification tasks: presence/absence of violence, patient status (victim, perpetrator and/or witness) and violence type (domestic, physical, sexual)	Pre-trained BioBERT Model	10-fold Cross-Validation; F1 score >0.89. Best F1 0.97 for sexual violence.
Analyzing the Impact of Domestic Violence on Social Media using Natural Language Processing	More and Francis	2021	English-Language Reddit posts, Tweets and news articles discussing domestic abuse	Unsupervised Topic Modelling	Latent Dirichlet Allocation (LDA) and other unsupervised topic models	Topic coherence - results not discussed
Utilizing Text Mining, Data Linkage and Deep Learning in Police and Health Records to Predict Future Offenses in Family and Domestic Violence.	Karystianis et al.	2021	Police-attended domestic violence event narratives from New South Wales, Australia. This data was linked with data about mental health diagnosis, age at diagnosis, episode start and end data, from the New South Wales Ministry of Health.	Multi-class time series prediction model (Predict probability of future offense in three categories: physical, non-physical, Apprehended Domestic Violence Order (ADVO) breach).	Transformer model with BERT embeddings	Best accuracy was 0.69 for predicting *ADVO breach* in multi-class classification
A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data	Rodriguez and Storer	2020	English-Language Tweets tagged with #WhyIstayed or #WhyILeft	Unsupervised Topic Modelling	Structural Topic Model with 65 topics	Held-out Likelihood, Residuals, Semantic Coherence, and Maximised Lower Bound used to evaluate goodness of fit for clustering - 65 topics identified as best fit according to these metrics
Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter	Xue et al.	2019	Tweets about Domestic Violence	Unsupervised Topic Modelling	LDA with 20 topics	Rate of Perplexity Change (RPC) used to evaluate best number of topics
Sexual Harassment Story Classification and Key Information Identification	Liu et al.	2019	Safecity English-language narratives of sexual harassment	Entity Recognition (four entity types: harasser, time, location, trigger words);	Convolutional Neural Network (CNN)	Accuracy: 0.92 for Entity Recognition. Highest accuracy for multi-class classification was

Name of Study	Authors	Year	Dataset	Model Task	(Most Successful) Model Type	Evaluation Outcome (of best performing model)
			incidents (originally published by Karlekar and Bansal)	Multi-class classification (5 story types)		0.97 achieved by a CNN model for "Time of Day"; for "type of harasser" this was 0.93
Apply event extraction techniques to the judicial field	Li et al.	2019	Chinese-Language litigation texts from divorce proceedings	Event Extraction (13 types of event in divorce cases, one of which is an event of domestic violence)	Combined architecture consisting of dictionary methods and 2x Long Short-Term Memory (LSTM) and Conditional Random Field (CRF) models.	10-fold Cross Validation; F1 = 0.84
Understanding the silence of sexual harassment victims through the #WhyIDidntReport movement	Garrett and Hassan	2019	Tweets using the #WhyIDidntReport hashtag from US cities	Multi-class classification (8 different reasons for not reporting sexual violence e.g. shame, feeling hopeless, lack of information, protecting perpetrator etc.)	Support Vector Machine (SVM)	F1 ranged from 0.47 - 0.78 across different classes.
Corpus-driven insights into the discourse of women survivors of Intimate Partner Violence	Sanchez-Moya	2017	British domestic violence charity online forum posts	Linguistic Analysis	Linguistic Inquiry and Word Count. (LIWC)	N/A
An analysis of domestic abuse discourse on reddit	Schrading et al.	2015	Reddit posts discussing abuse and not relevant to abuse	Binary Classification (About abuse / not about abuse)	Linear SVM	0.92 Accuracy was the best performance achieved on a train/test set of submissions concatenated with their top-scoring comments. Also tested 'in the wild' on Reddit data from general 'relationship advice' forums.
Indirect Identification of Perinatal Psychosocial Risks from Natural Language	Allen et al.	2021	Diary entries of pregnant women	Binary Classification (IPV - non-IPV)	LASSO Regression with sentiment, topic modelling and LIWC as features	Regression R^2, Area Under Curve = 0.08, 0.75 (test set)
Online Social Support for Intimate Partner Violence Victims in China: Quantitative and Automatic Content Analysis	Chu et al.	2021	Chinese-Language posts from Baidu Teiba group about intimate partner violence	Multi-class Classification (Emotional Support, Informational Support or None)	Logistic Regression	Classification Accuracy = 0.94; F1 = 0.56 (test set)
Deep Learning for Multi-Class Identification From Domestic Violence Online Posts	Subramani et al.	2019	Facebook posts relating to domestic violence	Multi-class Classification (Type of Support)	CNN and LSTM-type models; SVM with Term Frequency-Inverse Document Frequency (TF-IDF) vectors as features	Classification Accuracy = 0.91; F-Measure = 0.91 (3-fold CV)

Name of Study	Authors	Year	Dataset	Model Task	(Most Successful) Model Type	Evaluation Outcome (of best performing model)
Domestic Violence Crisis Identification From Facebook Posts Based on Deep Learning	Subramani et al.	2018	Facebook posts about "domestic violence and domestic abuse"	Binary Classification (Crisis - non-crisis)	LSTM-type models	Classification Accuracy = 0.94 (10 fold CV)
Child Abuse and Domestic Abuse: Content and Feature Analysis from Social Media Disclosures	Subramani et al.	2018	Facebook posts about domestic abuse and child abuse	Binary Classification (Abuse-related, General)	Decision Tree with Pscyho-linguistic Features (LIWC)	Accuracy = 0.95; F1 = 0.94 (10 fold CV)
Intent Classification Using Feature Sets for Domestic Violence Discourse on Social Media	Subramani et al.	2017	Facebook posts from four Domestic Violence related Facebook groups	Binary Classification (Abuse – advice or support)	SVM with 15 selected LIWC features	Classification Accuracy = 0.97 (10 fold CV)
Quantitative Methods for Analyzing Intimate Partner Violence in Microblogs: Observational Study	Homan et al.	2020	#WhyIStayed and #WhyILeft Tweets	Binary Classification (#WhyIStayed - #WhyILeft)	Radial Basis Function SVM	0.78 Accuracy
The hidden pandemic of family violence during COVID-19: Unsupervised learning of tweets	Xue et al.	2020	COVID-19 and Domestic Violence related Tweets in English	Unsupervised Learning (Cluster Analysis)	Latent Dirichlet Allocation (LDA)	N/A - unsupervised classification
Automated Identification of Domestic Violence in Written Child Welfare Records: Leveraging Text Mining and Machine Learning to Enhance Social Work Research and Evaluation	Victor et al.	2021	Records of referrals of child maltreatment from Michigan, USA	Binary Classification (DV - non-DV)	K Nearest Neighbours (KNN) Model, k=30	Classification Accuracy = 0.91 (5 fold CV)
Knowledge Discovery in Databases - Amsterdam-Amstelland Police Force Research Project	Poelmans et al. Collected Studies	2008-2013	Dutch police reports of violent events from the Amsterdam-Amstelland Police Force	Knowledge Discovery	Rule-based model developed using Formal Concept Analysis in combination with a Emergent Self Organising Maps visualisation; SVM and KNN with ESOM-enhanced inputs	Accuracy = 0.91 for incoming police reports; Accuracy = 0.89 for existing police reports; Accuracy = 0.95 for existing reports
Text-Mining Police Reports from New South Wales Project	Karystianis et al. Collected Studies	2019-2022	Police-attended domestic violence event narratives from New South Wales, Australia	Information Retrieval. Identify mentions of: mental health conditions, abuse types and injury type. Generate descriptive statistics on demographics	Rule-based dictionary model, IBM SPSS Statistics	Precision/F1: 0.9/0.89 for types of abuse, 0.85/0.86 for victim injuries

3.4.1. Models and Techniques

3.4.1.1. Model Task

The studies constructed models for a wide variety of tasks. The most common task was a binary classification task involving some form of 'abuse' vs. 'not-abuse' categorisation (e.g. Victor et al. 2021 (DV, non-DV); Subramani et al. 2018 (Abuse related, General); Allen et al. 2021 (IPV, non-IPV); Schrading et al. 2015 (About abuse, not about abuse)). Three studies were concerned with information retrieval tasks rather than classification (Karystianis et al. Collected Studies, Li et al. 2019, Liu et al. 2019). Several studies framed post or narrative types as a classification problem (Homan et al. 2020; Subramani et al. 2017, 2018a, 2019; Chu et al. 2021; Garrett and Hassan 2019). Of the unsupervised studies, most were concerned with unsupervised topic modelling (Xue et al. 2020; Xue et al. 2019; Rodriguez and Storer 2020; More and Francis 2021). Finally, only two studies were concerned with identifying different types of abuse: Karystianis et al. Collected Studies and Botelle et al. 2022. No studies explored different sub-types of abuse, such as financial abuse, psychological abuse or coercive control, in detail.

3.4.1.2. Supervised Techniques.

Supervised techniques are those that are developed using a *labelled* dataset – a dataset where each instance has been annotated (labelled) with an outcome or category (for example, each Tweet in a Twitter corpus is manually labelled with either 'about domestic abuse' or 'not about abuse'). These existing annotations can be used as a benchmark to evaluate automatic text mining methods, which makes supervised techniques a popular choice. The majority (n=16) of the included studies used some kind of supervised approach. Supervised techniques are also the basis for many ML models. Supervised ML models 'learn' patterns from the labelled dataset to create an accurate model that can then be applied to new, unseen data

(Alpaydin, 2020). This is an extremely convenient way to extend classification tasks to a dataset that is much larger than could be annotated by hand (Botelle et al., 2022).

As described in Chapter 2, there are two broad types of Supervised ML models: Traditional and Deep Learning models. Traditional models, such as Support Vector Machines (SVMs), K-Nearest Neighbours (KNN), LASSO Regression, and Decision Trees (DTs) iteratively try to find the best fit for the boundaries between one or more classes[4] in a high dimensional space - a process commonly referred to as model training. In over a third of the included studies (n=8) a Traditional Supervised ML model was the main, or most successful, approach (Allen, Davis, & Krishnamurti, 2021; Chu et al., 2021; Garrett & Hassan, 2019; Homan et al., 2020; Schrading, Alm, Ptucha, & Homan, 2015a; S. Subramani, Vu, & Wang, 2017; S Subramani, Wang, Islam, Ulhaq, & O'Connor, 2018; Victor et al., 2021), with SVMs being the most common successful model (Garrett & Hassan, 2019; Homan et al., 2020; Schrading et al., 2015a; S. Subramani et al., 2017).

Deep Learning models were used as the main approach in six studies (Botelle et al., 2022; Karystianis, Cabral, Han, Poon, & Butler, 2021; Li et al., 2019; Y. Liu et al., 2019; S. Subramani et al., 2019; S. Subramani et al., 2018), often using a traditional ML model as a baseline. Deep Learning models are very large networks of decision nodes – known as neural networks – which discover extremely complex multi-dimensional relationships between input and output (Alpaydin, 2020). Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are two broad families of Deep Learning models (S. Subramani et al., 2019).

[4] In ML literature, *class* refers to an outcome, category or label that a model is trying to optimise for. For example, if a model was being built to automatically categorise (or *classify)* Electronic Health Records as to whether or not they contained a mention of domestic abuse, the two *classes* would be “abuse present” and “abuse absent”.

Transformer based models, such as BERT (Devlin et al., 2018), are very large deep models that have already learnt a statistical representation of a language (most commonly, English) from huge amount of data. For instance, the original BERT model was trained on a corpus of books and Wikipedia entries of over 3 billion words (Devlin et al., 2018). Since these pre-trained models have a wide 'understanding' of language already, they are very adaptable to new tasks, even those where there is little data available. One included study used BioBERT (Lee et al., 2020), an adaptation of the original BERT model specifically suited for biomedical text mining tasks, to identify instances of DA in Electronic Health Records (Botelle et al., 2022).

Deep Learning models often achieve better results than Traditional ML in complex tasks (Botelle et al., 2022; S. Subramani et al., 2018). However, their drawback is their high level of opacity, which explains why they are frequently being referred to as 'black boxes'. Processes like feature ablation[5] (Karystianis et al., 2021) and dimensionality reduction[6] (S. Subramani et al., 2019) can help to visualise and understand the most important factors in the decision of a model. Additionally, recent advances in the domain of explainable machine learning have resulted in tools such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016a) which can be used to provide insight into the decision-making mechanisms of Deep Learning models. Nonetheless, their results can still prove difficult to interpret (Karystianis et al., 2021).

[5] Feature ablation refers to removing one or more features and observing the change in model performance in order to understand how different features affect the decision of a model

[6] Dimensionality Reduction refers to the process of transforming a very high-dimensional space in a space with lower dimensions, whilst preserving important characteristics of the data. This can help to visualise clusters or decision boundaries within a complex model

The remaining two studies which used a supervised approach used rule-based models to automatically classify data, using existing labels to test the accuracy of their rules (Karystianis et al., 2022; J Poelmans, Van Hulle, et al., 2011). Hand-crafted rule-based models have the advantage of being very transparent and efficient in comparison to ML models. It is probably not a coincidence that the two studies which used this approach were both actively working with police forces, who are likely to value transparency highly. Rule-based models performed very well in both studies (0.89 F1-score for abuse types (Karystianis et al., 2022); Accuracy >0.89 for identifying domestic violence in police reports (J Poelmans, Van Hulle, et al., 2011)). This suggests that they should not be overlooked in favour of more modern but complex tools such as Deep Learning models.

3.4.1.3. Unsupervised Techniques.

Six studies used unsupervised topic modelling or exploration as their primary approach (More & Francis, 2021; Rodriguez & Storer, 2020; Sanchez-Moya, 2017; Xu, Zeng, Tai, & Hao, 2022; Xue et al., 2020; Xue et al., 2019). Here 'unsupervised' is used to mean that a dataset has no labels or annotations - it is simply a collection of instances of raw text data (for example, a collection of Tweets *without* any categories or labels assigned to each Tweet)

3.4.1.4. Unsupervised Clustering.

Four of the six studies used Unsupervised Machine Learning (Unsupervised ML) models, which analyse the latent structure of a text corpus to identify related clusters, or *topics*, in a process called *topic modelling*. The most common topic modelling approach was Latent Dirichlet Allocation (LDA), used in three studies (More & Francis, 2021; Xue et al., 2020; Xue et al., 2019), whilst the other study used Structural Topic Modelling (STM) (Rodriguez & Storer, 2020).

3.4.1.5. Unsupervised Exploratory Approaches

Two studies used forms of exploratory data analysis as their primary method of investigating text data. Xu et al. (Xu et al., 2022) deployed a custom rule-based approach to *sentiment analysis.* The latter describes the practice of analysing texts according to their positive or negative emotional tone.

Sanchez-Moya (2017) used Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001), a computational tool for linguistic analysis. This technique was also used in four other studies as an addition, or an input into, more complex models (Allen et al., 2021; Rodriguez & Storer, 2020; S Subramani et al., 2018). LIWC is a *dictionary-based* method, in that it counts the number of words in a text which belong to a series of dictionaries of words from particular linguistic categories (e.g. positive affect, negative affect, biological processes, analytical thinking, emotional tone) (Sanchez-Moya, 2017). Dictionary-based methods are a simple but powerful instrument than can be very efficient, and used across multiple studies, once the hurdle of creating the initial dictionary has been passed. Other included studies created their own dictionaries of DA-related terms (Adily et al., 2021; Li et al., 2019; J. Poelmans, Elzinga, Viaene, & Dedene, 2009).

3.4.1.6. Technologies.

Matlab, R and Python were mentioned most often as technologies used in the studies, reflecting their popularity for data science applications. At least seven studies mentioned using Python (Chu et al., 2021; Garrett & Hassan, 2019; Homan et al., 2020; More & Francis, 2021; Schrading et al., 2015a; Xu et al., 2022; Xue et al., 2019), although many studies did not report any specific technology or programming language used.

Table 2

Datasets Used in Included Studies

Dataset	Authors	Year	Type	Source	Perspective	Geography	Language	Size	Labels
Chinese social media comments about Yuya IPV disclosure	Xu et al.	2022	Social Media posts	Chinese Social Media	Mix	Chinese-speaking world	Chinese	34,350	N/A
Electronic Health Records (EHRs) from the South London and Maudsley NHS Foundation Trust	Botelle et al.	2022	Electronic Health Records	South London and Maudsley NHS Foundation Trust	3rd Party	UK	English	5,282	Manual - 2 Student Reviewers, weekly labelling meetings
English-Language Reddit posts, Tweets and news articles discussing domestic abuse	More and Francis	2021	Social Media posts	Reddit, Twitter and News	Mix	English-speaking world	English	Unspecified	N/A
Police-attended domestic violence event narratives from New South Wales, Australia	Karystianis et al. Collected Studies	2019-2022	Police Reports / Mental Health Data	New South Wales Police Force, Australia	3rd Party	Australia	English	492,393	Existing - internal police officer classification process
English-Language Tweets tagged with #WhyIstayed or #WhyILeft	Rodriguez and Storer	2020	Social Media posts	Twitter	Mix	English-speaking world	English	3060	Automatic - According to Hashtags
Tweets about Domestic Violence	Xue et al.	2019	Social Media posts	Twitter	Mix	English-speaking world	English	322,863	N/A
Safecity English-language narratives of sexual harassment incidents (originally published by Karlekar and Bansal)	Liu et al.	2019	Social Media posts	Safecity social media platform	1st Person Reported	English-speaking world	English	9,892	N/A
Chinese-Language litigation texts from divorce proceedings	Li et al.	2019	Litigation Texts	Divorce proceedings	3rd Party	China	Chinese	3,100	Manual - Events annotated according to Beginning-Inside-Outside (BIO) method
Tweets using the #WhyIDidntReport hashtag from US cities	Garrett and Hassan	2019	Social Media posts	Twitter	Mix	USA	English	37,526	N/A

Dataset	Authors	Year	Type	Source	Perspective	Geography	Language	Size	Labels
British domestic violence charity online forum posts	Sanchez-Moya	2017	Social Media posts	British Domestic Violence Charity Forum	Mix	UK	English	472	Automatic - According to forum topic
Reddit posts discussing abuse and not relevant to abuse	Schrading et al.	2015	Social Media posts	Reddit	Mix	English-speaking world	English	370,410	Automatic - According to forum topic
Diary entries of pregnant women	Allen et al.	2021	Diary Entries	Diary Entries from pregnant women	1st Person Reported	USA	English	309	Automatic - according to psychometric measures
Chinese-Language posts from Baidu Teiba group about intimate partner violence	Chu et al.	2021	Social Media posts	Baidu Teiba	Mix	Chinese-speaking world	Chinese	4,800	Manual - 2 Student Reviewers
Facebook posts relating to domestic violence	Subramani et al.	2019	Social Media posts	Facebook	Mix	English-speaking world	English	1,654	Manual - 2 Student Reviewers
Facebook posts about "domestic violence and domestic abuse"	Subramani et al.	2018	Social Media posts	Facebook	Mix	English-speaking world	English	2,060	Manual - Multiple reviewers
Facebook posts about domestic abuse and child abuse	Subramani et al.	2018	Social Media posts	Facebook	Mix	English-speaking world	English	4,239	Automatic - According to search term
Facebook posts from four domestic violence related Facebook groups	Subramani et al.	2017	Social Media posts	Facebook	Mix	English-speaking world	English	8,856	Manual - 2 Student Reviewers
#WhyIStayed and #WhyILeft Tweets	Homan et al.	2020	Social Media posts	Twitter	Mix	English-speaking world	English	17,534	Manual - 4 student reviewers
COVID-19 and domestic violence related Tweets in English	Xue et al.	2020	Social Media posts	Twitter	Mix	English-speaking world	English	1,015,874	N/A
Records of referrals of child maltreatment from Michigan, USA	Victor et al.	2021	Case Summaries	Records of child maltreatment in Michigan, USA	3rd Party	USA	English	75,809	Manual - 4 student reviewers
Dutch police reports of violent events from the Amsterdam-Amstelland Police Force	Poelmans et al. Collected Studies	2008 -2013	Police Reports	Amsterdam-Amstelland Police Force	3rd Party	The Netherlands	Dutch	9,552	Existing - internal police officer classification process

3.4.2. Datasets

3.4.2.1. Source.

The majority of the datasets used in the included studies were sourced from social media (n=15) with the remainder coming from police forces (n=3), health services (n=1), litigation proceedings (n=1), children's social workers (n=1), and a single study which directly recruited participants (n=1). A summary of the datasets can be found in Table 2.

As expected from a search conducted in English, the majority of datasets (n=18) are in English, with the others being in Chinese (n=3) and Dutch (n=1). Of those datasets sourced from a particular locality (e.g. police data), the US, UK, Australia, China and the Netherlands are represented. Datasets are notably missing from other countries where English is widely spoken, such as Canada, India, Pakistan, South Africa or Nigeria. Around a quarter of the datasets (n=6) describe abuse from the perspective of a 3rd party reporting on the abuse (e.g. a police officer or healthcare professional). Conversely, a small number (n=2) describe abuse from the perspective of the victim-survivor narrating their own experience(s). The remaining datasets (n=14) contain a mix of perspectives (e.g., social media groups where some posts are from the victim-survivor perspective and some are from 3rd parties describing abuse which happened to someone else, or offering support). No datasets explore either text written from the perspective of a perpetrator, or direct evidence of abuse in text (e.g. abusive text messages).

3.4.2.2. Size.

The size of the datasets varies considerably, from 309 diary entries (Allen et al., 2021) to over 1 million unique Tweets (Xue et al., 2020). The size of each text within a dataset also varies, from a single Tweet (Homan et al., 2020) to entire litigation texts (Li et al., 2019) or case summaries (Victor et al., 2021). Of the datasets used for supervised ML tasks, the average size was 73,847 instances.

3.4.2.3. *Labelling process.*

Data labelling is often a time consuming and costly part of computational text mining, which can discourage research from taking place in new areas. In addition, data labelling has a direct impact on the outcome of classification models, since any bias or inaccuracies in the labelling process are likely to be picked up and replicated by the model (Bechmann & Zevenbergen, 2019; Dignum, 2017). For this reason, accurate and transparent labelling is of paramount importance, especially in sensitive research.

Most datasets were labelled by supervised student reviewers. However, some datasets took advantage of existing properties of the data to create labels – for example, by using hashtags applied to tweets (Homan et al., 2020), participant surveys administered alongside the collection of text data (Allen et al., 2021), or police assigned labels collected during the incident reporting process (J Poelmans, Van Hulle, et al., 2011). Such techniques can significantly reduce the time and cost burden for researchers, and show the benefit of trying to find label-type properties within existing data.

3.4.3. Evaluation

3.4.3.1. *Test and Train set.*

A *test set* is a portion of the dataset that is set aside during model development, and subsequently used to evaluate the algorithm's final performance on held-out data. Leaving part of the data out during model development helps avoid *overfitting*, where models learn the statistical characteristics of a dataset "too well", in a way that means their results don't generalise to other data (Arango et al., 2019). For small datasets, a mechanism called *k-fold*

Cross Validation (k-fold CV) is often used to evaluate a model's performance, in combination with or instead of a separate test set. This involves separating the data into *k* different segments. The model is then allowed to see all but one of these segments when it is training, and after training has finished, the left-out segment is used to test the model. The process is then repeated k times, each time leaving out a different segment. The results of these k times are then averaged to give an overall evaluation metric.

3.4.3.2. Evaluation Metrics.

All studies using supervised techniques were evaluated using a test set or k-fold CV. *Accuracy* and *F1 score* were the most common metrics used to report how well the model performed at correctly categorising the texts in the test set. Accuracy refers to the overall percentage of instances which were correctly classified. The F1 score metric is an alternative metric which balances *Precision* (also known as specificity, or true negative rate) and *Recall* (also known as sensitivity, or true negative rate). The F1 score is useful in situations where one class is much larger than another – in this case, Accuracy scores can be unhelpfully biased towards the dominant class (Hugo Rosa et al., 2019).

However, comparison of models across different datasets using reported metrics should be done cautiously, since much of the performance of a model depends on the data it was trained on. Some datasets simply have too much overlap between the characteristics of different classes, making it difficult for a model to distinguish between them.

Taking into account these comments on the limitations of metrics, there is a very wide range of accuracies in the studies, from 0.69 (which would usually be considered too low to be used in any practical application) (Karystianis et al., 2021) to 0.97 (as good of a performance as can reasonably be expected from most models) (Botelle et al., 2022). There was no single type of model or technique which performed well across the studies. This reflects the variability of model tasks within the studies, and demonstrates the importance of choosing the right model for the task in question.

3.4.3.3. Unsupervised Evaluation.

Evaluation of the studies which used unsupervised approaches was much more variable, reflecting the difficulties in evaluating unsupervised methods more broadly (Zhao et al., 2015). Some unsupervised studies did not include any explicit evaluation of their technique (Xu et al., 2022) or were using tools developed and tested in previous research (such as LIWC (Sanchez-Moya, 2017)). Other studies which used unsupervised topic modelling attempted to evaluate the optimal number of topics, using methods such as Rate of Perplexity Change (RPC) (Xue et al., 2019).

3.5. Discussion

Overall, the N=22 studies showcase different models and techniques which can be used for DA research, as well as a variety of datasets and evaluation mechanisms. This section provides a more detailed discussion of the reviewed studies, focusing on the quality of current work, lessons learned for future research, and ethical issues raised by using computational methods to research DA.

Table 3

Quality of Studies

Author	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Victor et al. 2021	✓	✓	✓	✓			✓			✓	✓		✓	✓		✓	✓	✓	✓	✓	
Botelle et al. 2022		✓					✓			✓						✓	✓	✓	✓	✓	✓
More and Francis 2021				✓	✓	✓	✓	✓	✓	✓	✓	✓					N/A	N/A	N/A	N/A	N/A
Schrading et al. 2015	✓	✓		✓		✓	✓			✓	✓	✓					✓	✓	✓	✓	✓
Karystianis et al. Col. Studies	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓			✓		✓	✓	✓	✓
Xu et al 2022	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	N/A	N/A	N/A	N/A	N/A
Karystianis et al. 2021		✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	
Rodriguez and Storer 2020		✓		✓	✓	✓	✓			✓	✓	✓	✓	✓		✓	N/A	N/A	N/A	N/A	N/A
Xue et al. 2019				✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	N/A	N/A	N/A	N/A	N/A
Liu et al. 2019		✓				✓	✓	✓		✓	✓	✓						✓	✓	✓	
Li et al. 2019		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				✓	✓				
Garrett and Hassan 2019	✓	✓		✓			✓										✓	✓		✓	✓
Sanchez-Moya 2017	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	N/A	N/A	N/A	N/A	N/A
Allen et al. 2021	✓	✓		✓		✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓		✓	
Subramani et al. 2017					✓	✓	✓	✓	✓	✓	✓		✓				✓	✓	✓		
Subramani, Wang, Islam et al. 2018	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓					✓	✓			
Subramani, Wang, Vu and Li. 2018	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓		
Subramani et al. 2019	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓			✓	✓	✓		
Chu et al. 2021	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓	✓	
Homan et al. 2020		✓		✓	✓	✓	✓						✓			✓	✓	✓	✓	✓	
Xue et al. 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	N/A	N/A	N/A	N/A	N/A
Poelmans et al. Col. Studies	✓	✓	✓	✓			✓			✓	✓		✓				✓	✓		✓	✓

3.5.1. Quality of Studies

Results from the 21-criteria Quality Assessment are reported in Table 3. The quality of studies varied considerably across the included works. This reflects the innovative nature of this new, interdisciplinary area. Furthermore, there are also not yet clear guidelines about how to use text mining methodologies in social science research. In addition, challenges arise when attempting to assess quality across such a heterogeneous set of studies. For example, some papers did not report any pre-processing steps (Criteria 6) since this is not useful in Deep Learning architectures (S. Subramani et al., 2018). Other studies did not report demographic characteristics of their dataset (Criteria 3) due to ethical concerns about collecting personal identifiers alongside sensitive data (Rodriguez & Storer, 2020; Xue et al., 2019).

3.5.2. Lessons for Future Research

Examining aspects of the included studies could offer lessons for future research, particularly regarding the definition of violence, open source computational research, and overall study design. These issues are discussed in more depth below.

3.5.2.1. Definition of Violence.

The definition of violence is mentioned in just over half the studies (n=13), but many do not define DA at all, or very briefly reference a definition from another entity, such as the WHO (Chu et al., 2021). Studies tend to discuss the definition of violence in most detail when examining the dataset labelling process for supervised techniques. Labelling data often highlights conflicting definitions between annotators and necessitates a more in-depth description of what constitutes violence (Botelle et al., 2022; J. Poelmans, Elzinga, Viaene, & Dedene, 2009). However, most studies did not comment in depth, or at all, on the labelling process of their datasets, or report on discussions about the definition of abuse between annotators. Overall, future researchers should carefully consider how text mining models will capture and encode a specific definition of DA, and include more detail about the process of coming to an agreement around the concept of DA when annotating datasets.

3.5.2.2. Open Source.

Unfortunately, no projects in the study reported that their code or data was *open source*. The latter describes a trend in computer science to make code and data available freely online, to facilitate collaborators wishing to build similar applications. Two projects mentioned that their dataset would be made available upon request (Botelle et al., 2022; Xu et al., 2022). This is perhaps unsurprising when it comes to datasets, given the sensitive nature of the data involved. However, future work could consider making source code available for other researchers, to encourage knowledge-sharing within this emerging field.

3.5.2.3. Study Design.

In general, future projects could consider a number of factors in study design. Firstly, researchers may reflect where novel data can be sourced, and whether data from multiple sources can be joined-up for additional insight (Karystianis et al., 2021). Secondly, once a model has been developed, researchers could consider deploying or testing it in an active service-provision environment. For example, research projects from Poelmans et al. (2013) and Karystianis et al. (2022) successfully worked with police forces to implement knowledge-discovery techniques within their day-to-day operations, and models revealed edge cases of abuse that the police had previously missed (Hwang et al., 2020; J. Poelmans, Elzinga, Viaene, & Dedene, 2009). A project to detect DA in Electronic Health Records is now live on systems of an NHS trust in the UK (Botelle et al., 2022). The transformer-based model used in the study is able to detect sexual and physical violence, as well as whether the patient was the perpetrator or victim-survivor of violence.

Moreover, when designing methodologies, researchers must consider more than just the choice of model. Rule-based, Traditional ML, Deep Learning and Unsupervised approaches all performed well in different included studies, demonstrating that the context and appropriateness of a model is more important than its type. The importance of initial data exploration and feature selection should not be ignored, as these processes (referred to as *feature engineering*[7]) significantly increase the quality of outcomes. For example, Subramani et al. (2017) did not use the raw text, but instead the outcome of LIWC (see Unsupervised Exploratory Approaches, above), as the input to their ML model (S. Subramani et al., 2017). Finally, several studies highlighted the importance of mixed methods in their research, and the significance of pairing quantitative methods with qualitative insights (Rodriguez & Storer, 2020; Victor et al., 2021).

3.5.3. Ethical Concerns

In general, too little attention was paid to ethics across the studies, with only six publications including an explicit ethical discussion. However, a large number (n=14) of studies do mention limitations of their work or discuss appropriate contexts for model use.

For example, Victor et al. indicate that whilst their model performs well enough to be used for generating accurate descriptive statistics about domestic violence in a dataset of child welfare case summaries, it would be inappropriate for use in decision making about individual cases (Victor et al., 2021). They highlight the importance of qualitative analysis when using ML methods in an interdisciplinary context, giving three examples of how qualitative analysis can enrich ML research in this domain: understanding the data-generating

[7] Feature engineering is the set of steps that transform raw data into numeric values that are usable by ML models. These numbers are features that represent each instance/sample (e.g. an abusive sentence or a non abusive one) and are used as inputs by the ML models

mechanism, its context, content and what inferences can reasonably be made; understanding outliers and misclassifications in order to improve the model; and applying insights from the model to “help standardize or build consistency in how domestic violence is assessed and documented” (Victor et al., 2021).

Allen et al. comment on the lack of diversity in their sample, which contained mostly white participants, and how this could exacerbate existing problems of under-reporting depression and DA in other racial groups (Allen et al., 2021). However, very few studies commented on the demographic representativeness of their dataset and whether or not downstream applications built on their models risked excluding certain groups.

Given the recent emphasis within ML communities on ethical principles of accountability, responsibility and transparency (Dignum, 2017; Floridi et al., 2018), future work must take more of a focus on discussing the foundational ethical questions raised by this kind of research. Researchers might consider following ethical guidelines for ML such as those proposed by the Association of Internet Researchers (Bechmann & Zevenbergen, 2019).

The consequences of ignoring such ethical discussions are significant: At their worst, ML models could contribute to the invalidation and minimisation of different experiences of abuse, for example by classifying an instance as “not abuse” and leading to a victim-survivor not receiving services or justice after having experienced great harm (Blackwell, Dimond, Schoenebeck, & Lampe, 2017). The “scientific” interpretation of a model risks being taken as more “legitimate” or “accurate” than a victim-survivors “emotional” interpretation. Victim-survivors of DA have experienced situations in which they have had their opinions and experiences repeatedly invalidated, belittled, denied and manipulated (Stark, 2009).Researchers must, thus, be aware of the potential mis-use of their research to extend this denial of the victim-survivor's reality. Models are representations of reality, but they are not reality themselves, and the way ML research is conducted and presented should reflect this understanding.

3.5.4. Limitations

The systematic review is subject to several limitations. Firstly, since the search strategy only included academic literature, it is possible that important grey literature may have been missed. Secondly, the search terms included other types of violence such as "family violence" and "sexual violence", aiming to capture all definitions of violence that may include DA. Some of the reviewed studies may therefore have included incidents of other types of abuse in their data. Finally, the Quality Assessment criteria used in the review were developed by combining multiple existing methods and were not thoroughly evaluated on different types of studies outside this review. They should therefore not be used as a ranking mechanism or to draw concrete conclusions about the quality of individual studies.

3.6. Conclusion

Twenty-two studies which used computational text mining to investigate DA were identified through a systematic literature review of eight academic databases. The studies included datasets from social media, police forces, a healthcare provider, and social work and legal settings. A variety of supervised and unsupervised text mining techniques were used on these datasets for tasks which included detecting the presence or absence of DA as well as identifying abuse types, extracting entities and events, or understanding themes. Some studies commented on the ethics or application of their findings, but future research could include more in-depth discussion of these. Potential areas for future research include sourcing datasets from other geographies and types of organisations, as well as further research into sub-types of abuse. In particular, the review identified a lack of research using computational methods to examine psychological abuse. Furthermore, datasets presented in the paper suffered from a lack of explanation as to their labelling process and to the definition of abuse used when constructing their datasets.

Chapter 4: Pursuing Data Explainability in a Labelled Dataset of Reddit posts for Machine Learning Classification of Types of Psychological Abuse

4.1 Introduction

The systematic review presented in Chapter 3 identified a gap in the literature concerning the use of computational methods to study psychological abuse. Whilst a few of the included studies mentioned “psychological abuse” or similar terms, none sought to use machine learning to identify specific psychologically abusive behaviours. Furthermore, Chapter 3 highlighted that existing literature on using ML to study DA often glosses over the process of labelling, what definitions were used for ‘abuse’ and how disagreements were resolved. Glossing over this process may result in more satisfying results for publication, since the model appears to match the ground truth well – but if the ground truth itself is not robust, explainable, or widely agreed upon, then the results of the classifier are less meaningful.

4.1.1. Psychological Abuse

As introduced in Chapter 2, psychological abuse is a contested term, often used to refer to patterns of behaviour which cause psychological harm in the context of an intimate relationship (SafeLives, 2019). Psychological abuse is still not as well understood as physical abuse, partly because it is more difficult to define (Dokkedahl et al., 2019; Follingstad, 2009; Lagdon et al., 2014), and because it is highly contextual, often tailored by the perpetrator to target the victim-survivor in a way that remains unnoticeable to an observer (Crown Prosecution Service, 2017; Stark, 2009).

In this context, studying psychological abuse in an organic online space, such as Reddit (Reddit.com), gives the opportunity to understand how victim-survivors report psychological abuse at a large scale, from their own direct perspective, without the constraints of specific survey or interview questions. Reddit was chosen as a suitable platform for researching discussions of abuse online for multiple reasons: Firstly, the design of the Reddit platform - absent of identifying profile pictures, and with most users using pseudonymous usernames - encourages a culture of anonymity, which may facilitate the discussion of sensitive topics, such as mental health (De Choudhury & De, 2014) and DA (Sivagurunathan, Walton, Packham, Booth, & MacDermid, 2021; Trinh Ha, D'Silva, Chen, Koyutürk, & Karakurt, 2022). Secondly, since Reddit is a publicly accessible online forum without a paywall, it is not necessary to create an account to access posts, and it is practical to scrape data from reddit using the Pushshift API (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020). In addition, most discussion on Reddit is primarily in text format (not images or videos). This makes Reddit particularly suitable for scraping-based research. Reddit discussions are also organised into Subreddits around different topics, which makes the curation of appropriate data easier than on other platforms such as X (Proferes, Jones, Gilbert, Fiesler, & Zimmer, 2021). Finally, Reddit has a number of active communities discussing domestic abuse.

Other possible platforms to study included Facebook (S. Subramani et al., 2017), X (Homan et al., 2020) or forums from the websites of specific domestic abuse organisations. The latter presented barriers in terms of negotiating appropriate data sharing agreements with domestic abuse organisations, which was not feasible for the current research. Since the research aimed to collect personal narratives, the micro-blogging format of X was deemed to be unlikely to contain sufficient information to identify psychological abuse. Facebook allows longer posts, and has a very large user base which might enable a collection of data from a wider variety of users. However, since Facebook requires a user to be logged in to view some pages, and many Facebook groups containing sensitive discussions are invite-only, automated data scraping of Facebook is more challenging than Reddit. For these reasons, Reddit was chosen as the most suitable data source for this research.

The creation of a dataset of reports of psychological abuse on Reddit was motivated by the desire to understand how common reports of psychological abuse are on the platform, and what types of psychologically abusive behaviour are reported. The large volume of posts on Reddit make machine learning classifiers extremely valuable as a research method in answering these questions.

However, due to the contested and complex definition of psychological abuse, it is important that the process of creating these ML classifiers, and their training dataset, be conducted in a transparent way. This chapter therefore brings together research from the DA and responsible Artificial Intelligence (AI) fields, to present the process of creating an 'explainable' dataset of different types of psychological abuse.

4.1.2. Explainable Artificial Intelligence

Much work around responsible Artificial Intelligence (AI) has focused on the *explainability* of black-box models through post-hoc illumination of their decision-making mechanisms (Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016b; Rudin et al., 2022; Xie et al., 2020). However, less attention has been paid to the *explainability* of the *data* that has been used to build these models. This is despite increasing recognition that a machine learning (ML) model can only ever be as good as the data it was trained on, and that models will inevitably reflect inaccuracies, biases and misunderstandings contained within their training data (S. Barocas & Selbst, 2016; Whang, Roh, Song, & Lee, 2023).

Some AI researchers have begun to emphasise *data-centric* machine learning, the practice of increasing model performance through improving the quantity, and sometimes quality, of training data (DeepLearning AI, 2021; Whang et al., 2023), but work in this field has tended to focus on augmenting and enlarging datasets without thorough investigation, or reporting, of more tricky aspects of data such as transparency, bias and the annotation process.

Some scholars have started to critically examine the process of creating training data and to voice concerns about common data annotation practices (Aroyo & Welty, 2015; Kapania et al., 2023; Röttger et al., 2021). This chapter furthers this discussion through use of the expression *data explainability,* a term starting to appear online (Pichaiah, 2023) and in some academic literature (McDermid, Jia, Porter, & Habli, 2021).

Since *data explainability* is a new term, it is necessary to define here what is meant by explainable data and how this chapter draws on existing literature to attempt to create an explainable dataset. To do so this chapter will first give some background on the existing literature around data annotation.

4.1.2.1. Data Annotation for Machine Learning

In order to train a supervised ML model, it is necessary to create a ground truth dataset for the model to learn from. This often involves the process of data annotation, also known as labelling, in which a subset of available data is manually annotated with the correct output which the model should attempt to emulate for each data instance.

For example, to train an ML model to recognise positive or negative movie reviews, it would first be necessary to annotate a dataset of film reviews and label each one as "positive" or “negative” depending on their sentiment. What seems like a simple task on the surface veils a minefield of decisions – what if a review seems “neutral”? What if it contains irony or contradiction that is difficult to interpret as “positive” or “negative”? What if one annotator is sure that a review is “positive”, but a second annotator from a different cultural background interprets the same review as “negative”?

This example attempts to illustrate some of the problems that inevitably crop up when labelling data. Indeed, data annotation is often viewed as a time consuming, laborious and frustrating process which, according to one practitioner, “tends to alternate between mind-numbingly boring and excruciatingly painful” (Muller et al., 2021). Data annotation is recognised to be “fraught with problems” (Kulesza et al., 2014), arising from the fact that the abstract “concept” underlying the labels can vary between even highly specialised individuals, be influenced by bias, and change over time (Kulesza et al., 2014).

In an attempt to reduce the effect of human subjectivity, one commonly employed strategy is the use of multiple annotators (Kulesza et al., 2014). The use of multiple annotators rests on the idea of the “wisdom of the crowd”, and that it is possible to come to a consensus about labels by somehow averaging between the perspectives of multiple people – for example, by taking the majority vote (Davani, Díaz, & Prabhakaran, 2022).

However, existing work has highlighted the subjective nature of truth and the difficulty of flattening multi-annotator data into a single ‘truth’ (Aroyo & Welty, 2015). A particular pain-point in data annotation with multiple annotators is the resolution of instances where different annotators disagree on the correct label (Muller et al., 2021). Kapania et al. (2023) point out that disagreement is often seen as “undesirable, impeding production of ‘high quality data’”

(Kapania et al., 2023) and any evidence or discussion of disagreement is often not included in the final dataset, despite the fact that this disagreement might be useful to demonstrate that a phenomenon is contested or elicits diverse viewpoints amongst different groups. Similarly, Aroyo and Welty (2015) identify seven common myths about human annotation of data, including "there is one truth" and "disagreement is bad" (Aroyo & Welty, 2015).

To make matters worse, ML datasets are often notably lacking documentation or in-depth discussion of how they are labelled, rendering the "design process" of the data "less and less visible" (Muller et al., 2021). Notably, when interviewed, those involved in data labelling processes often *did* acknowledge that engineering is a process of trade-offs and that labels were a "series of approximations" subject to "resource limitations, such as available staff and available time" (Muller et al., 2021). However, despite this acknowledgement of the messy process involved in labelling data, these "series of approximations" and "resource limitations" are rarely mentioned in research chapters or documentation which accompany datasets.

Against this backdrop, several scholars have proposed innovative methods for increasing transparency, diversity and replicability in ML datasets. It is these collective efforts that are referred to in this chapter using the term *data explainability* – any processes engaged in to make the source and design process of annotated data more transparent, robust and replicable.

Existing work in this area includes work by Rottger et al (2021), who propose two paradigms for labelling datasets – a prescriptive and a descriptive paradigm, where the prescriptive paradigm explicitly aims to train annotators to capture a single version of a concept, whilst the descriptive paradigm aims to capture a multiplicity of opinions amongst annotators.

Röttger et al. propose that dataset creators should understand, identify and make explicit which paradigm they wish to work in when labelling datasets (Röttger et al., 2021). Similarly, Kapania et al. (2023) suggest using "annotation tools and processes that support a multiplicity of voices" (Kapania et al., 2023).

Muller et al. propose the 'preservation of labelling histories' for a particular instance in order to preserve the metadata about who applied a label, when, in what context, and how the label changed over time (Muller et al., 2021). Prabhakaran, Davani and Diaz (2021) suggest that datasets should be released alongside disaggregated annotator labels and socio-demographic information of annotators as well as documenting how annotators were recruited (Prabhakaran et al., 2021).

There is therefore an emerging body of research which addresses how to make data more , transparent, robust and replicable. However, despite this existing literature, very few datasets have been published which actually attempt to address these concerns and make their data fully explainable. This gap in the literature makes it difficult for dataset creators to apply general principles of *data explainability* even if they want to, due to a lack of best practice or practical guidelines. The creation of more explainable datasets is of paramount importance to the responsible AI mission – without transparent and explainable data, it is not possible to train truly transparent and explainable machine learning models (S. Barocas & Selbst, 2016). Furthermore, since benchmark machine learning datasets are often used for decades once published (Yang, Qinami, Fei-Fei, Deng, & Russakovsky, 2020), they create a kind of data 'lock-in' (Crootof, 2019) - so if data isn't made more explainable now, it may be considerably more difficult to make it so in the future.

4.1.3. Research Questions

This chapter therefore presents a manually labelled text dataset that has been created with *data explainability* at its heart. The dataset is created in a three-stage process which 1) draws from domain specific literature to create an annotation scheme which is grounded in existing research 2) refines the annotation scheme through iterative expert discussion and 3) contains both individual and aggregated labels from all annotators in the dataset, including where annotators disagree. This is accompanied by a discussion of lessons learned from this process which other researchers might draw from in creating their own explainable datasets. Overall, the chapter is motivated by three research questions:

1. What existing definitions of psychologically abusive behaviour exist in literature, and how can we incorporate these into an annotation scheme?
2. To what extent can a panel of expert annotators agree on an annotation scheme for psychological abuse?
3. How can we use a transparent annotation process to make the final dataset *explainable*?

4.1.4. Chapter Outline

The remainder of the chapter is structured as follows: 1) Methodology: The methodology of this study is described, including the scraping of data from Reddit and the three-step *data explainability* process used to create the dataset. 2) Results: The annotation scheme is presented and the dataset is described through aggregate statistics, including measures Inter Annotator Agreement. 3) Discussion: The *data explainability* process and the resulting dataset are discussed, including lessons learned, limitations, and potential directions for future work. 4) Concluding remarks.

4.2. Method

The method consists of a *data explainability* process with three steps, as illustrated in Figure 4, using data scraped from Reddit forums about domestic abuse:

1. A collection of existing definitions of psychological abuse and associated behaviours are collated from literature and form the basis of an annotation scheme of different types of psychological abuse.
2. The annotation scheme is iteratively refined through a series of expert discussionss.
3. The full dataset of Reddit posts is labelled using the annotation scheme, inter-annotator agreement is calculated, and the dataset includes individual annotations as well as different types of aggregated annotations.

The following section initially describes the process of scraping data from Reddit, and subsequently describes each of the three *data explainability* steps in turn.

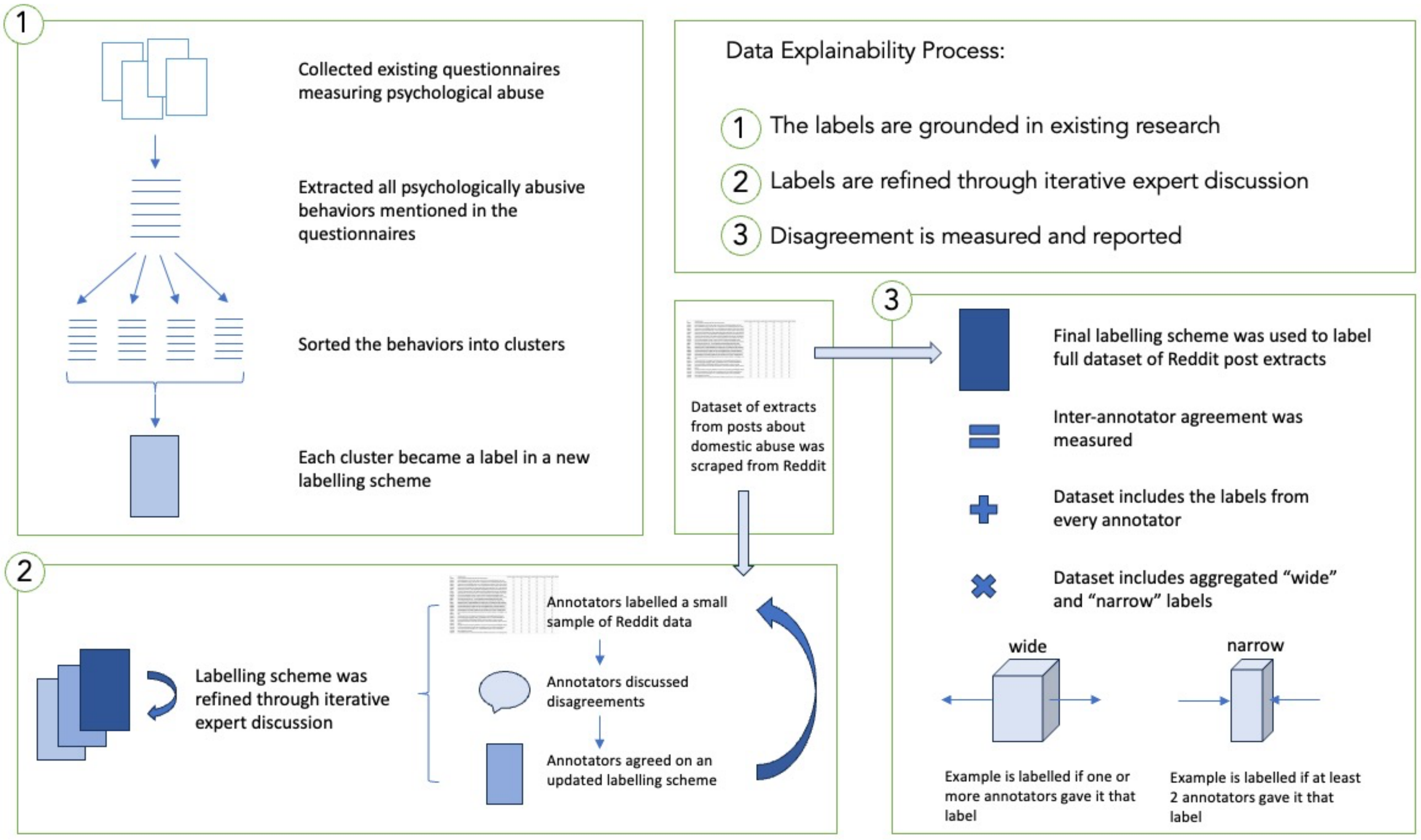
4.2.1. Data Scraping from Reddit

An initial qualitative analysis of discussions of psychological abuse on Reddit was performed by searching for “psychological abuse” and related terms (“emotional abuse”, “coercive control”, “mental abuse”, “narcissistic abuse”, “abusive relationship”) on Reddit.com in early 2022. This initial search revealed a number of Subreddits containing posts relating to psychological abuse. Post types included narrative accounts of experiencing DA, as well as help-seeking posts from both victim-survivors and friends or family, posts offering advice or encouragement, and a few posts appearing to be from the perspective of perpetrators of abuse.

Further qualitative analysis of the top posts in each subreddit indicated three subreddits of high relevance for the current study based on the following criteria: they mostly contained discussions about emotional or psychological violence (rather than other types of violence e.g. physical violence, or tangential topics such as e.g. experiences of PTSD); they mostly discussed abuse in the context of a current or former intimate partner relationship; they contained a high proportion of posts detailing victim-survivor stories of abuse, rather than e.g. help seeking behaviour or perpetrator perspectives; and finally, the total number of members was of an appropriate size for the practical purposes of this study (between 20 and 75k members).

The resulting three subreddits (r/abusiverelationships, r/domesticviolence, r/emotionalabuse) were scraped using the Pushshift Reddit API (Baumgartner et al., 2020). To limit the volume of data, only submissions posted during 2021 were collected, since this was the last complete year at the data collection stage of the project. This resulted in a raw dataset of 59,106 submissions. Eliminating duplicate, empty and deleted posts reduced this to 46,519 submissions.

Initial bi-gram and tri-gram analysis showed some unexpected results, which revealed that a few prolific members of the forum were posting very frequently and skewing linguistic analysis. Qualitative analysis indicated that the initial post from a poster was most likely to provide background information on their story, making it the most useful post for our analysis. Therefore, only the first post from each username was kept in the final dataset. This resulted in an edited dataset of 28,630 posts. Initial labelling experiments involved labelling entire posts, however some posts consisted of several pages of text and were too long to label clearly. To reduce ambiguity and decrease labelling time, texts were shortened by randomly sampling a 400-character extract from each post. Finally, to reduce the number of posts to a manageable size for manual annotation, 2000 posts were randomly sampled to create a dataset for labelling.



Data Explainability Process:

1 The labels are grounded in existing research

2 Labels are refined through iterative expert discussion

3 Disagreement is measured and reported

1

Collected existing questionnaires measuring psychological abuse

Extracted all psychologically abusive behaviors mentioned in the questionnaires

Sorted the behaviors into clusters

Each cluster became a label in a new labelling scheme

Dataset of extracts from posts about domestic abuse was scraped from Reddit

2

Labelling scheme was refined through iterative expert discussion

Annotators labelled a small sample of Reddit data

Annotators discussed disagreements

Annotators agreed on an updated labelling scheme

3

Final labelling scheme was used to label full dataset of Reddit post extracts

Inter-annotator agreement was measured

Dataset includes the labels from every annotator

Dataset includes aggregated “wide” and “narrow” labels

Example is labelled if one or more annotators gave it that label

Example is labelled if at least 2 annotators gave it that label

Figure 4 – *Data Explainability Process Diagram*

4.2.2. Data Explainability 1: Grounding in Existing Literature

The first stage of our *data explainability* process aimed to develop a draft annotation scheme for different types of psychological abuse that was grounded in existing research on this topic. The aim of this stage was to make the epistemological process of conceptualising labels more transparent and replicable. An external researcher could attempt to re-create this process following the descriptions contained below. To the knowledge of the author, the grounding of the data annotation scheme in a literature review of existing domain knowledge represents an innovation in the field and has not previously been trialled in other research.

Firstly, a literature review was conducting to identify a) existing psychometric measures of psychological abuse and b) existing frameworks for understanding types of psychological abuse. This literature was then used to create a database of examples of psychologically abusive *behaviours* and an initial annotation scheme to categories these behaviours into different *types*.

4.2.2.1. Psychometric Measures of Psychological Abuse

The first part of the literature review aimed to identify psychometric measures of psychological abuse that are already in use by practitioners and researchers. Psychometric measures of psychological abuse are scales or questionnaires used by DA advocates, psychologists, police and other organisations to measure the perpetration or victimisation of psychological abuse. These measures tend to offer the subject a series of statements and ask them how strongly they agree with the statement, or how often the statement applies to them. In this way, collating these statements and identifying the behaviours they mention across the different measures can give a picture of what is commonly considered psychological abuse by practitioners in the field.

The systematic review protocol of Dokkedahl et al. (Dokkedahl et al., 2019) served as the jumping off point for identifying psychometric measures of psychological abuse. In total, 18 measures were identified from Dokkedahl et al.'s review and other literature as a) measuring perpetration or victimisation of psychological or emotional abuse b) containing descriptions of abusive behaviour (rather than, for example, describing only the impact of behaviour) c) having at least one unique behaviour not already mentioned in another measure. The 18 measures can be found in Table 4.

For each of these 18 measures, every behaviour mentioned in the questionnaire or scale was extracted into a database, with similar behaviours from different questionnaires merged into a single item. This process resulted in a database of 251 potentially psychologically abusive behaviours, such as: "criticism or regulation of household duties or childcare", "controlling or managing walking or posture", "saying victim is crazy or irrational", "saying hurtful actions will help make victim a better person", "yelling or screaming", "suicide or self-harm threats or attempts", "keep victim from having time to themselves" and so on.

These behaviours ranged from behaviour that might seem relatively benign, when taken out of context (such as, for example, "bringing up the past to hurt" or "overreaction to incidents", "swearing", "contradictory statements or rules") to behaviours that in and of themselves would be considered illegal or abusive ("Threats to kill", "forced victim to steal", "rape", "having victim followed or following victim"). This demonstrates the importance of context and a pattern of behaviour when identifying psychological abuse (Crown Prosecution

Service, 2017; SafeLives, 2019), as mentioned in the Background section. It is therefore important to note that whilst this database includes behaviours that are often part of a pattern of abuse, individual behaviours in and of themselves are not necessary ‘abusive’.

Table 4

Measures of Psychological Abuse

Name	Author(s)	Year Published	Size	Source/Reference
Psychological Abuse of Women by Spouses and Live-in Lovers	Hoffman	1984	71 statements in 22 categories	(Hoffman, 1984)
Abusive Behavior Inventory	Shepard & Campbell	1992	30-item scale	(Thompson et al., 2006)
Partner Abuse Scale - Non-Physical	Hudson	1992	25-item scale	(Thompson et al., 2006)
Measure of Wife Abuse	Rodenburg & Fantuzzo	1993	60-item scale	(Thompson et al., 2006)
Index of Spouse Abuse	Campbell et al.	1994	30-item scale	(D. W. Campbell, Campbell, King, Parker, & Ryan, 1994)
Safe Dates - Psychological Abuse Victimisation	Foshee et al.	1998	14-item scale	(Thompson et al., 2006)
Composite Abuse Scale	Hegarty, Sheehan, & Schonfeld	1999	30-item scale	(Thompson et al., 2006)
Index of Psychological Abuse	Sullivan et al.	1999	33-item scale	(Thompson et al., 2006)
Multidimensional Measure of Emotional Abuse	Murphy et al.	1999	28-item scale	(Thompson et al., 2006)
Profile of Psychological Abuse	Sackett & Saunders	1999	21-item scale	(Thompson et al., 2006)
Psychological Maltreatment of Women Inventory - Female Version	Tolman	1999	58-item scale	(Tolman, 1989)
Subtle and Overt Psychological Abuse Scale (SOPAS)	Marshall	1999	Extract of 39 items form full scale used	(Marshall, 1999)
Controlling Behaviours Scale - Revised (CBS-R)	Graham-Kevan & Archer	2003	24-item scale	(Graham-Kevan & Archer, 2003)
Revised Danger Assessment	Campbell et al.	2004	20-item questionnaire	(J. C. Campbell, Webster, & Glass, 2009)
Follingstad Psychological Agression Scale	Follingstad et al.	2005	17-item scale	(Follingstad et al., 2005)
Coercion in Intimate Partner Relationship Scale	Dutton et al.	2005	48-item scale	(Dutton, Goodman, & Schmidt, 2005)
Measure on Psychologically Abusive Behaviours	Follingstad	2010	14-item scale	(Follingstad, 2011)
Escala de Abuso Psicológico Aplicado en la Pareja (EAPA-P) (English version)	Porrúa-García et al.	2016	47-item scale	(Porrúa García et al., 2016)

4.2.2.2. Frameworks of Psychological Abuse

Evidently, 251 behaviours is too large a number to form a coherent annotation scheme; therefore it was necessary to cluster these behaviours into a small number of *types* of behaviour (e.g. isolating behaviour, threatening behaviour), which would form the labels in an annotation scheme. To do this, the literature review looked at existing frameworks of psychological abuse to understand how DA researchers had conceptualised *types* of psychological abuse. Three frameworks were identified, as shown in Table 5. The frameworks had some overlap – for example, all three include isolation as a category – but there were also differences, for example surveillance/monitoring is a category in the Stark and Marshall frameworks, but not in the Duluth Power and Control Wheel.

Table 5

Existing Frameworks for Categorising Psychological Abuse

Framework Name	Author	Year	Categories
Coercive Control Framework	Stark	2013	Threat; Regulation; Surveillance; Isolation; Degradation; Deprivation; Exploitation
Power and Control Wheel	Duluth Domestic Abuse Intervention Program	1984 - present	Using coercion and threats; Using intimidation; Using emotional abuse; Using economic abuse; Using male privilege; Using children; Minimizing, denying and blaming; Using isolation
Subtle and Over Psychological Abuse Scale	Marshall	1999	Dominate; Indifference; Monitor; Discredit; Undermine; Discount; Isolate

Each of these frameworks was then experimented with as an annotation scheme, by sorting the 251 behaviours into the different types from each framework. Table 6 demonstrates this process, showing a sample of the 251 behaviours and which types they fit into according to the three different frameworks.

Table 6

Examples of Abusive Behaviours Categorised Using Different Frameworks

Behaviour	Power And Control Wheel	SOPAS Themes	Stark (2013)
Saying victim is crazy or irrational	Using emotional abuse	discredit	Coercion: Threats
Threats to leave	Using coercion and threats	dominate	Coercion: Threats
Silent treatment or ignoring, refusing to talk to victim	Using emotional abuse	indifference	Coercion: Threats
Did not allow or pressured not to work	Using economic abuse	isolate	Control: Deprivation & Exploitation
Issuing orders & demanding obedience	Using male privilege	dominate	Control: Regulation
Micromanagement of cleaning, cooking or household chores	Using Intimidation, Using male privilege	dominate, monitor	Control: Regulation
Blame victim for own poor/violent behaviour, for all problems in relationship	Minimizing, denying and blaming	undermine	Coercion: Degradation
Pressure victim to perform non-consensual sexual acts	Using Intimidation, Using coercion and threats	Sexual aggression	Coercion: Degradation
Hostile, rude, suspicious or jealous interactions with victim's friends	Using isolation	isolate	Control: Isolation

Through this process, Stark's Coercive Control framework was identified as being easiest to use to label behaviours, with most behaviours fitting clearly into one category in this framework. Stark's framework was therefore used as an initial annotation scheme in the first of a series of labelling discussions. Furthermore, the 251 behaviours in the database were subsequently used as examples during the expert discussions, helping to add detail and clarity to the annotation process.

4.2.3. Data Explainability 2: Refining Through Iterative Expert Discussion

Following the literature review, a four-member annotation team came together for a series of discussions during April and May 2022 to iteratively develop the proposed annotation scheme. This stage of the *data explainability* process followed best practice as proposed in, for example, (Muller et al., 2021) on refining an annotation scheme through expert discussion. This iterative approach has been used in other machine learning work on domestic abuse to solidify conceptualisations and definitions of abuse amongst a team of annotators (Botelle et al., 2022).

4.2.3.1. Annotation Team

The annotation team consisted of four researchers within the DA field who were recruited through professional networks. According to recommendations from Prabhakaran et al., we report demographic characteristics of the annotators here (Prabhakaran et al., 2021): All annotators were female and university educated. Three annotators were white and one of Asian background. Three out of four annotators had English as a first language. The annotators had different cultural backgrounds: Two annotators were from the UK, one from the US, and one from China. All annotators had previously worked on material concerning DA before starting the study, with one annotator having a significant professional background in the DA sector. Two members of the annotation team had previously worked on a different study which also involved labelling a dataset about DA. The annotation team was therefore habituated to reading distressing material, and had a baseline familiarity with behaviours and patterns often seen in DA.

4.2.3.2. Expert Discussions

A series of discussions were conducted with the annotation team in order to iteratively refine the annotation scheme. An initial annotation scheme based on the existing literature was proposed and iteratively refined by testing it on a small sample. Annotators would then discuss disagreements or uncertainties with each other.

Data Explainability 3: Calculating Disagreement and Aggregating Labels

4.2.4. Labelling of Full Dataset

Following the labelling discussions, a dataset of 2000 posts was labelled by all annotators in the combination showed in Table 7. At least two annotators labelled every example in the dataset.

In order to understand the level of agreement between annotators, Inter-Annotator Agreement (IAA) (also known as Inter-Rater Reliability (IRR)) was calculated for all annotators on the first 500 examples using the Fleiss Kappa statistic (Fleiss, 1971; Landis & Koch, 1977). There are multiple measures of IAA, the choice of which depends on the number of annotators and the type of annotation (e.g. categorical, nominal, ordinal) (Gisev, Bell, & Chen, 2013). Fleiss Kappa was chosen above other measures of agreement because the data to be labelled was categorical, because annotations were sourced from more than two annotators, and because not all annotators labelled all data (Gisev et al., 2013). Fleiss Kappa is an adaptation of Cohen's Kappa, a statistic which measures the level of agreement above that which could be expected by chance, but which is only usable in situations with exactly two coders who both annotate all examples. Fleiss Kappa is an adapted measure than can be used in situations with multiple annotators where not all annotators sample all data (Fleiss, 1971; Hallgren, 2012).

Table 7

Examples Labelled by Each Annotator

	Annotator A	**Annotator B**	**Annotator C**	**Annotator D**
1-500	Labelled	Labelled	Labelled	Labelled
501 - 1000	Labelled	Labelled		
1001 - 1500	Labelled		Labelled	
1501 - 2000	Labelled			Labelled

4.2.4.1. *Aggregated Labels: Narrow and Wide Definitions*

Following both the recommendations suggested by Prabhakaran et al. (Prabhakaran et al., 2021) and proceeding in the descriptive paradigm proposed by Rottger et al. (Röttger et al., 2021) which aims to capture a multiplicity of beliefs, the dataset includes disaggregated labels from all four individual annotators. In this way, the dataset does not hide the disagreement amongst annotators. The "design process" of the dataset is therefore made more "visible" (Muller et al., 2021).

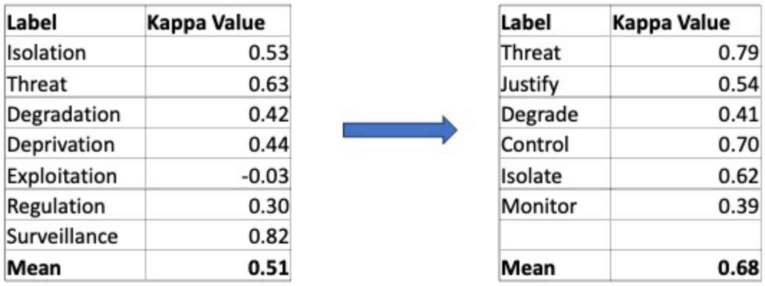
In addition, in order to facilitate easier training of a machine learning model, whilst still capturing the disagreement between annotators, the dataset includes two types of aggregated labels: Aggregated (Wide) and Aggregated (Narrow) capture the broadest and narrowest definitions respectively of each psychologically abusive behaviour according to our panel of annotators. In the "wide" labels, an instance is labelled as positive for a category if *any one* *annotator* had labelled that instance as positive for that category. In the “narrow” dataset, an instance is labelled as positive in a category if *at least 2 annotators agreed* on an example and both labelled that instance as positive for that category. Note that the labels are not mutually exclusive (see full explanation of the annotation scheme in the Results section), so each example could be given zero, one, or multiple labels.

4.3. Results

The results of the study are presented in two parts: first, the annotation scheme which resulted from the literature review and expert discussions is presented and discussed; and second, the resulting dataset is presented including aggregate statistics and measures of Inter Annotator Agreement.

4.3.1. Annotation Scheme

The draft annotation scheme used for the first workshop was drawn from the literature review described above. This initial annotation scheme had seven labels taken from Stark’s conceptualisation of coercive control (Stark, 2013). Over the course of discussions, this evolved into a new, six-label annotation scheme that achieved a higher level of inter annotator agreement on a small sample of posts (see Figure 5).



Label	Kappa Value
Isolation	0.53
Threat	0.63
Degradation	0.42
Deprivation	0.44
Exploitation	-0.03
Regulation	0.30
Surveillance	0.82
Mean	**0.51**

Label	Kappa Value
Threat	0.79
Justify	0.54
Degrade	0.41
Control	0.70
Isolate	0.62
Monitor	0.39
Mean	**0.68**

Figure 5 – *Change in Annotation Scheme and Inter Annotator Agreement (Fleiss Kappa) over the Expert Discussions*

Note: Agreement shown in Figure 5 was calculated between two annotators since not all annotators were present at all discussions. The sample size was n=32 and n=50 at the start of discussions (left hand table) and end of discussions (right hand table) respectively

The final labels were as follows:

1. *Rules, Control and Micro-regulation*
2. *Justify, Minimize and Deny abuse*
3. *Threats, Intimidation and Punishment*
4. *Shaming, Degrading and Ignoring*
5. *Isolation*
6. *Surveillance, Monitoring and Harassment.*

The labels are *not mutually exclusive*, meaning a post could fall under none, one, or multiple labels. An extensive set of annotation guidelines were developed for the scheme including examples for each label. An extract from our annotation scheme can be seen in Table 8.

Table 8

Annotation Scheme resulting from the Expert Discussions

Label	Outcome for Victim	Selected Examples of Included Behaviours
Rules, Control and Micro-regulation	Cause victim to change behaviour	- *Constant criticism of victim's behaviour* - *Victim not allowed to disagree* - *Micro-regulation of victim's dress* - *Limiting, controlling or regulating food consumption* - *Sleep deprivation* - *Issuing orders and demanding obedience* - *Imposing decisions about children* - *Sabotaging or withholding birth control*
Justify, Minimize and Deny abuse	Cause victim to accept the abuse	- *Manipulating or hiding information to suit own interests* - *Denial of victim's perceptions* - *Get victim to apologise or feel guilty for something that wasn't their fault* - *Making light of the abuse or saying it didn't happen* - *Contradictory statements and erratic behaviour* - *Saying victim is crazy or irrational*
Threats, Intimidation and Punishment	Cause victim to feel scared or anxious	- *Swearing* - *Yelling or screaming* - *Verbal threats or warning* - *Threats to leave, have an affair* - *Threats to suicide or self harm* - *Threats involving children* - *Physical violence* - *Extreme irritability and mood changes* - *Destroying property* - *Displaying weapons* - *Disappearing with explanation* - *Sending harassing messages*
Shaming, Degrading and Ignoring	Cause the victim to feel ashamed, small, weak	- *Criticism or shame of ideas or proposals* - *Criticism of intelligence or work* - *Act like victim doesn't matter* - *Act like there is something wrong with victim mentally* - *Putting down physical appearance* - *Bringing up the past to hurt* - *Silent treatment or ignoring* - *Trying to discredit victim publicly*
Isolation	Cause victim to be cut off from support	- *Pressuring victim not to work* - *Restricting victim's use of car* - *Trying to turn friends or family against victim* - *Hostile, suspicious or jealous interactions with victim's friends* - *Requiring victim to be home a lot* - *Monitoring time and whereabouts* - *Moved far away from friends and family* - *Showing up unexpectedly from work* - *Insisting on coming along to meetings with family and friends*
Surveillance, Monitoring and Harassment	Cause victim to feel watched or trapped	- *Using video cameras to spy on victim's activities* - *Timing partner on the phone* - *Using children to pass messages* - *Monitoring time and whereabouts* - *Monitoring eating* - *Using IoT devices to monitor* - *Insisting on location sharing for 'safety'*

4.3.1.1. Overlap Between Labels

Some labels overlapped significantly with other labels, resulting in smaller "sub-clusters" of related behaviours that fell into more than one category. For example, behaviours that were labelled as both *Threat, Intimidation and Punishment* and *Rules, Control and Micro-regulation* were generally those that controlled or managed behaviours that the victim-survivor should be able to decide for themselves. Examples included controlling a partner's weight, controlling how they dressed, spoke, or washed themselves, withholding food or medication, controlling their consumption of media, or controlling access to birth control or other medical treatment. Control of such activities could be used as both a rule which the victim has to follow (e.g. if they don't dress in a certain way, the partner will humiliate them and put them down), or a punishment for transgression (e.g. if they do something the perpetrator doesn't like, they will withhold their medication), meaning that these behaviours fall into both aforementioned clusters.

Similarly, a sub-cluster of behaviours that were both *Threat, Intimidation and Punishment* and *Degrading, Shaming and Ignoring* consisted of behaviours that used humiliation as a form of punishment – for example, ridiculing the partner in front of friends and family, calling them names, or making the partner do something degrading such as begging - as a retaliation for breaking one of the perpetrator's rules. This is similar but distinct from another sub-cluster of behaviours consisting of those that were both isolating and controlling – which mainly consisted of rules concerning the interactions a partner has with friends, family or other social support.

Isolation from social support weakens the victim's ability to leave the relationship or even to see that the relationship is unhealthy, since they no longer have access to outside perspective – as Stark describes, "Controllers isolate their partners to prevent disclosure, instil dependence, express exclusive possession, monopolize their skills and resources, and keep them from getting help or support" [Pg.27](Stark, 2013). Stark's framework of abuse conceptualises all isolating behaviours as controlling, whereas our taxonomy distinguishes some isolating behaviours which are more degrading, threatening or shaming than controlling – such as locking the victim in the house, humiliating them in public, or getting them to believe that other people don't care about them, for example.

4.3.1.2. Victim vs. Perpetrator Perspective

One of the problems arising when studying self-reported abuse is the narrator is portraying themselves in a subjective way. In Reddit forums discussing abuse, whilst most posters are writing about being the victim of an abusive relationship, a few posters are discussing their own perpetration of abusive tactics, usually to express guilt or confusion about their behaviour. During initial labelling discussions, a number of posts appeared in the sample where it was difficult to distinguish between whether the poster was experiencing or perpetrating abusive behaviour – often these overlapped, and the poster wanted to discuss whether their negative behaviour towards their partner was justified by the partner's initial behaviour towards them.

Through workshop discussions it was concluded that to avoid ambiguity, each post would be treated as if the poster is the victim-survivor of abuse, not the perpetrator. This would allow for the creation of a dataset that explores reported psychological abuse *victimisation*. Perpetrated behaviours would therefore not be counted in the sample (e.g. "he yelled at me multiple times" would be labelled as abuse, but "I yelled at her multiple times" would not be included). This is not because perpetration is not important or a fascinating subject for future work, but in order to avoid ambiguity and annotator assumptions in identifying the poster as either "perpetrator" or "victim".

4.3.1.3. Examples of Disagreement

Despite a very detailed annotation scheme there were still examples in the discussions where it was very difficult for all four annotators to agree on the correct annotation for a post. For example:

> "*He made lists of sexual acts that he wants from me. He threatens me and says he'll never speak to me again if I don't do the things on the list. He tells me I have to do what he says because he 'gave me another chance', but I don't know what that means – he cheated on me before, so why is it 'my chance' now?" - Paraphrased Post*

For all annotators this post extract was a clear example of both *Threats, Intimidation & Punishment* (the perpetrator threatening the poster that he'll never speak to them again) and *Rules, control & Micro-regulation* (giving the poster a list of sexual acts they are required to perform). Annotators also agreed that the post didn't mention *Isolation* or *Surveillance, Monitoring & Harassment***.** However, annotators disagreed as to whether this post was an example of *Justify, Minimize & Deny Abuse* or *Shaming, Degrading & Ignoring* behaviour. "*He tells me I have to do what he says because he 'gave me another chance'*" could be seen as a justification for the abuse by the perpetrator, but not all annotators agreed on this. Furthermore, cheating on the poster and then telling them they '*gave them another chance*' might also be considered degrading or shaming, since it could make the victim feel ashamed and small, but also isn't explicitly using shaming or degrading language towards the poster. Annotators disagreed on these latter two labels for this post.

> "*Last week he sneaked into my garden to watch me. I saw him and had panic attack, I was so frightened. His mother has a terminal illness. Anytime he accused me of*

> *cheating, or called me horrible names, he used to apologise and use his sick mother as an excuse to get me to go back to him." - Paraphrased Post*

Annotators agreed that this post was a clear example of *Surveillance, Monitoring & Harassment* (the perpetrator sneaking into the garden to watch the victim). Annotators also agreed that it was an example of *Shaming, Degrading & Ignoring* (calling the victim names) and *Justify, Minimize & Deny Abuse* (using a parent with a terminal illness as a justification for hurtful behaviour). However, annotators were split as to whether this was an example of *Threats, Intimidation & Punishment* – the poster mentions feeling very frightened, which suggests they felt intimidated and threatened by the perpetrator's stalking behaviour. At the same time, no explicitly intimidating behaviour or threats are mentioned.

These examples serve to highlight some of the reasons that annotators disagreed about labels and the kinds of discussions that were had during the discussions. They also demonstrate contextual nature of abusive behaviour and the difficulties in interpreting types of psychological abuse from narratives in online posts.

4.3.2. Dataset

The resulting annotation scheme was used to label the full dataset of 2000 posts as described in the Methods section, and Inter Annotator Agreement (IAA) was measured.

4.3.2.1. Inter Annotator Agreement

Table 9 shows the Fleiss Kappa statistic (Landis & Koch, 1977) calculated over the first 500 examples for all annotators and each combination of three annotators. The Fleiss Kappa statistic ranges between -1 and 1, where 1 indicates perfect agreement and -1 indicates perfect disagreement. A score over 0.6 indicates substantial agreement, 0.4-0.59 indicates moderate agreement, 0.21-0.4 indicates fair agreement, and below 0.2 indicates slight agreement (Landis & Koch, 1977).

Table 9

Inter Annotator Agreement: Fleiss Kappa

Annotators / Label	All	ABC	ABD	ACD	BCD
Rules	0.21	*0.16*	**0.25**	0.19	0.22
Justify	0.34	0.30	**0.43**	*0.27*	0.34
Threat	0.44	*0.37*	**0.61**	*0.37*	0.39
Shaming	0.31	0.26	**0.44**	0.27	*0.23*
Isolation	0.21	0.17	**0.31**	0.16	*0.15*
Surveillance	0.43	0.38	**0.58**	0.42	*0.34*

Note: Label names have been shortened for clarity. Bold indicates highest level of agreement for each label.

As can be seen from Table 9, *Threats, Intimidation & Punishment* is the least ambiguous of all the labels, achieving high agreement, and *Surveillance, Monitoring & Harassment* also achieves moderately high agreement. However, even after extensive discussions and iterative label development, two labels achieved very low agreement with every combination of annotators: *Rules, control & Micro-regulation* and *Isolation*. *Justify, Minimize & Deny Abuse* and *Shaming, Degrading & Ignoring* also achieved low agreement in all but the ABD annotator combination. The examples of disagreements given in the previous section, as well as the background on Psychological Abuse given at the start of this chapter, may help to explain this low level of agreement.

The highest level of agreement was achieved from the combination of Annotators ABD, which indicates that Annotator C had a higher level of disagreement with the other annotators and is potentially an outlier. It should be noted that Annotator C was the only Annotator who did not have English as a first language. This indicates that conceptions of psychological abuse may be very variable between different linguistic and cultural backgrounds.

4.3.2.2. Distribution of Labels

Table 10

Distribution of Labels within the Dataset

Label / Annotator	Rules, control & Micro-regulation	Justify, Minimize & Deny Abuse	Threats, Intimidation & Punishment	Shaming, Degrading & Ignoring	Isolation	Surveillance, Monitoring & Harassment
A	12.7%	19.15%	41.15%	20.35%	10%	5.45%
B	7.8%	7%	24.7%	9.9%	2.4%	3%
C	5.2%	4.2%	29.1%	5.9%	1.9%	3.5%
D	9.1%	9.5%	23.3%	21.6%	4.3%	4.7%
Aggregated (Narrow)	5.0%	6.4%	26.6%	10.3%	2.7%	3.1%
Aggregated (Wide)	17.8%	21.7%	46.6%	26.6%	11.3%	7.2%

Note: The table shows the percentage of posts that were given each label by each annotator (as a percentage of all the posts which that annotator labelled). Annotator A labelled all 2000 posts whilst annotators B, C and D only labelled 1000 posts each (see Table 7). A post can be labelled with none, one or multiple labels – the categories are not mutually exclusive. The

Aggregated (Narrow) dataset contains labels where at least two annotators agreed on that label. The Aggregated (Wide) dataset contains labels where any one annotator gave that label (See Method Section for full details) The wide and narrow datasets contain 2000 posts each.

Table 10 shows the distribution of labels in the dataset, as a percentage of the posts that each annotator labelled. For example, annotator C only labelled 1.9% of all the posts they annotated as containing *Isolation* behaviour. This is an absolute number of 19 posts, since Annotator C labelled 1000 posts in total. This indicates that for some labels, there were very few posts within the dataset that showed that type of psychological abuse.

4.3.2.3. Highly Imbalanced Data

Overall, looking at the low percentages in Table 10 indicates that the labels were too granular for the amount of data which was labelled, resulting in a highly imbalanced dataset (there are many more negative examples than positive examples for each label). The very small numbers of positive examples for some labels make it difficult to achieve high levels of inter-annotator agreement, or to get a clear picture of how exactly annotators disagreed.

To counter this, a keyword sampling approach was trialled, using dictionaries of words related to each label to try and increase the proportion of our sample belonging to each label (Botelle et al., 2022). However, keyword sampling did not result in a significant increase in the balance of several labels, and furthermore has drawbacks in potentially decreasing the efficacy of downstream models on unseen data. The ultimate solution would be to increase the amount of data labelled overall, but this was not possible due to time constraints – for context, each annotator spent between twenty and forty hours labelling data.

4.4.4.4. Most Common Types of Psychological Abuse

Despite disagreement between annotators, it is still possible to observe broad themes emerging from the distribution of labels, which help to further the conversation about reported psychological abuse. *Threatening, Intimidating and Punishing* behaviour was by far the most commonly reported form of abuse. *Surveillance, Monitoring and Harassment* behaviour was not commonly reported, which is surprising given that this is a frequently reported behaviour by victim-survivors (Office for National Statistics, 2018a). It is possible that victims of surveillance-type behaviours are less willing to post on public forums, given their experience of current or past surveillance.

Overall, it was observed during labelling that a very large number of posts reported experiencing some kind of psychologically abusive behaviour, but very few users mentioned engagement with any external service, such as the Police, shelters, social workers or DA charities. This indicates that many people experience psychological abuse without this ever coming to the attention of official services, which correlates with findings from other research (SafeLives, 2019).

4.4. Discussion

The Results of the study have shown the outcome of the *data explainability* process presented in the Methods section above. The dataset is accompanied with a transparent description of the annotation scheme, the annotation process, and the background of the annotators; furthermore, it embraces disagreement by including disaggregated labels from individual annotators. However, it was difficult for experts to reach high levels of agreement on different types of psychological abuse when labelling the dataset, despite extensive discussion of the annotation scheme during the expert discussions. For some types of psychological abuse, such as *Surveillance, Monitoring and Harassment*, the dataset was too small to see any meaningful results, because there were too few posts mentioning that type of behaviour.

The following section presents a discussion of these Results, focusing on the core question of this chapter: Is this data truly *explainable*? This is followed by a discussion of usability and replicability, which helps to frame the contributions and limitations of this work, and finally by a discussion of ethical considerations.

4.4.1. Is this data *explainable*?

The aim of the three stage *data explainability* process was to make the dataset discussed in this thesis more transparent, reliable and replicable – in other words, explainable. Firstly, grounding the data annotation scheme in existing literature served to illuminate the origin of the six-label categorisation of psychological abuse, making the process of coming up with the labels more robust, and replicable by a researcher who conducted a similar literature review to the one described in Section 4.2.2 above. Secondly, refining the annotation scheme through expert discussion aimed to make the annotated data more reliable, since data was labelled by multiple annotators and the iterative discussions allowed for the concepts underlying the labels to be refined and clarified. Finally, reporting disagreement between annotators made the data more transparent, since it is possible to identify in the dataset where annotators disagreed and where there are uncertainties about what constitutes different types of psychological abuse.

However, a limitation of this study is the lack of insight into the experiences and thought processes of the annotators. Future work could collect feedback from annotators in the form of “annotation diaries” or structured interviews conducted as part of the labelling process, which could help to illuminate annotator decision-making.

Furthermore, as well as measuring Inter Annotator Agreement to understand the disagreement between coders, it would be useful to understand *Intra*-Annotator Agreement to gain insight into the stability of annotator behaviour over time. As Belur et al. note in their work on IAA in systematic reviews, coder behaviour is often subject to both a learning effect (an increase in consistency of annotations as annotators learn over time) and a fatigue effect (a decrease in consistency of annotations as annotators become tired or frustrated with the task) (Belur, Tompson, Thornton, & Simon, 2021). Overall, it is unlikely that human annotators remain completely consistent in their application of the annotation scheme over multiple annotation sessions, at it would be useful to have more insight into how each annotator changed over time.

An additional limitation of the transparency of the dataset is that the annotators were all female, university educated and majority white. Future work in this area could seek to recruit a wider diversity of annotators.

4.4.2. Is this data *usable*?

A question that arose from the results of the *data explainability* process was – has explainability in this dataset come at the cost of usability? A more traditional annotation process, following majority voting and a prescriptive application of labels, rather than the collaborative development of labels from existing research, would probably have resulted in a higher level of agreement, and therefore a dataset that was on the surface easier to use for downstream machine learning applications. In addition, the dataset suffers from being too small, and if less time had been spent on labelling discussions and instead allocated to labelling data, the annotators might have been able to label more data and therefore increase the usefulness of this dataset.

However, the dataset presented here is usable in other ways. Firstly, the annotation scheme is easily adopted and re-used by future researchers looking at psychological abuse. In this regard, it is the first of its kind to the knowledge of the research team. Secondly, the annotation process in itself represents a contribution to DA research, since it represents a kind of qualitative analysis process similar to the process of coding in thematic analysis - for example, it is an interesting finding in its own right that surveillance behaviour is surprisingly rarely mentioned by Reddit posters. Finally, and most importantly, the dataset is still usable as training data for machine learning classifiers – but any resulting classifier would necessarily need to caveat its performance by indicating that it was trained on data with low levels of human agreement. If this encourages downstream research which moves away from presenting the predictive outcomes of machine learning models as a form of absolute truth, then this could be seen as a positive outcome.

4.4.3. Is the Data Explainability Process Replicable?

This chapter presented a three-stage process which aimed to make the resulting dataset explainable. One of the aims of the chapter was to demonstrate this process for future researchers seeking to create datasets in other domains. Broadly speaking, researchers could recreate this explainability process by: 1) conducting a literature review in their target field to identify existing definitions and frameworks for their subject matter, and adapting this existing work into an initial annotation scheme; 2) recruiting a panel of experts and engaging in iterative labelling experiments and discussions, whilst measuring IAA, to refine the initial annotation scheme; 3) measuring IAA over the full dataset and releasing the dataset along with disaggregated annotator labels, as well as a “wide” and “narrow” dataset, as described in the Results section. Some aspects of this process, particularly the literature review, the number and type of experts recruited, and the size and sampling method of the dataset, would need to be adapted to each particular domain.

Furthermore, lessons from this chapter indicate that particular attention should be paid to deciding the amount of data to label, which is directly proportional to the available time of the annotators. In addition, researchers might consider conducting a more formal workshop process in place of informal discussions, with annotators completing detailed annotation diaries or notes to provide insight into their decision-making process, that can then be reported alongside the dataset for further transparency. Finally, the demographic, linguistic and professional background of annotators is important to consider and discuss with the entire research team, since this is likely to have a significant impact on the outcome of annotations and the level of IAA.

4.4.4. Ethics

A number of ethical issues were taken into consideration and mitigated when conducting this research.

4.4.4.1. Researcher wellbeing

Repeatedly reading hundreds of stories of DA is an emotionally heavy and potentially distressing process for the researchers, and is also very time consuming. Our annotation team was working remotely, but we ensured at least one workshop was conducted in person, and regular online check-ins helped to ensure the psychological burden of this work was shared amongst the team during the labelling process. In addition, our annotators all had previous experience of working with data about DA, and so were aware of emotional challenges that can arise from this kind of research. Future work to create datasets about DA, and other similarly challenging topics, should consider the wellbeing of data annotators and how this will be protected as part of its ethical process.

4.4.4.2. Data scraping

Although all data used in this study derives from participants who knowingly posted on an anonymous but public online forum, they may not be aware that such data is commonly used for research purposes. Furthermore, individuals posting on the forums are likely to be victim-survivors of abuse, and are therefore a group with particular vulnerabilities and needs. All work using this data has been situated in an understanding and respect for the experiences of DA, and research team members with experience in DA advocacy were consulted on key decisions and research outcomes.

There is a risk that verbatim posts can be entered into search machines, used to identify usernames, and potentially recognise the individual concerned from their use of language or personal details without their awareness. To avoid this risk, the dataset is not being released publicly alongside this chapter, and instead should be requested directly from the researchers. In addition, any posts collected are not quoted verbatim – paraphrasing is used instead (Bechmann & Zevenbergen, 2019).

4.5. Conclusion

The initial step in the *data explainability* process presented in this chapter was to draw from existing research about psychological abuse by collecting and collating existing measures of psychological abuse. This led to the development of an annotation scheme with six non-mutually exclusive labels for different types of psychologically abusive behaviour. Whilst this six-label annotation scheme draws heavily on existing frameworks for psychological abuse, it is also slightly different in its conceptualisation, which represents a contribution to the study of psychological abuse. Such guidelines could be used for future research or annotation of a different dataset, since they contain detailed explanation and examples of what kinds of behaviour are contained within each label. Thus, the process of creating an annotation scheme through literature review and expert discussions, though ultimately aimed at building a machine learning model, also contributed to domain-specific knowledge for the DA field. This indicates that building machine learning datasets can contribute to furthering research in a particular domain, rather than playing a purely extractive, observational role.

Despite having a very detailed annotation scheme which was developed through a series of expert discussions, we did not manage to reach high levels of agreement for some labels. This reflects the fact that that psychological abuse is a contested phenomenon, even amongst experts, and appears in highly contextual and subtle ways that are difficult to fully systematise into clear labels.

This chapter has aimed to illustrate how difficult it can be in practice to solidify real-life, messy, complex concepts into a sufficiently concrete concept for a machine to learn from. In areas where there are limited human resources and it is difficult to train humans to make decisions on complex phenomena, it is helpful to imagine that machine learning models could do the hard work for us. This is particularly tempting when it comes to types of abusive crime, such as DA, bullying and hate speech, where manual labelling involves significant emotional labour from annotators. The chapter has built on existing work to argue that AI researchers and practitioners should bear in mind the limitations of creating training data around a concept that is still conceptually fluid and culturally debated. Ultimately, machine learning models cannot find the answer to every problem for which humans fail to find a solution. In areas as complex as DA, it would be a mistake to try to use AI to leapfrog the development of a clear societal understanding of what is and isn't appropriate or legal behaviour in human relationships. However, the *process* of improving *data explainability* can help to clearly illuminate the problems and debates in defining a contested phenomenon, and in this way contribute to domain knowledge.

Chapter 5: Automatic Recognition of Reports of Psychologically Abusive Behaviours using Machine Learning Classifiers

5.1. Intro

The systematic review in Chapter 3 identified a number of existing studies (n=14) which used machine learning text classification tools to identify different aspects of domestic abuse (DA) in their datasets. However, none of these studies applied such methods to detecting different types of psychological abuse. Chapter 4 therefore described the process of creating a dataset of different types of psychologically abusive behaviours, as reported by users on Reddit, and aimed to make the dataset 'explainable' through a transparent explanation of the annotation process.

A Machine Learning (ML) classifier that could automatically detect different types of psychological abusive behaviour would have a number of potential use cases. It would help researchers to identify psychological abuse in large datasets, which would increase understanding of how common psychological abuse is and how it presents in different populations. It could also potentially be used by law enforcement or DA advocacy organisations to identify individuals experiencing psychological abuse from existing databases, which could help to increase support for victim-survivors of psychological abuse. As described in Chapter 3, a number of existing studies have used supervised learning techniques to detect different aspects of DA in text, which indicates that ML methods can achieve useful results in this area.

5.1.1. Research Questions

This chapter therefore explores the application of ML methods to create an automatic classifier for types of psychological abuse, trained on the labelled dataset described in

Chapter 4. It also presents the results of using other computational methods to explore the dataset: particularly, an analysis of the most frequent n-grams per class; and unsupervised clustering of the entire dataset. This is motivated by the following Research Questions:

1. What insights can exploratory text mining give into the Reddit dataset of psychologically abusive behaviours?
2. Which ML models are most successful at classifying different types of psychologically abusive behaviour?
3. Which types of psychologically abusive behaviour are do ML models classify most successfully?

5.1.2. Chapter Outline

The remainder of the chapter is structured as follows: first, the Methods section describes the dataset, the exploratory analysis and pre-processing methods, and the classification models which were trained for the study; secondly, the Results section shows the results of the exploratory data analysis and the classifiers following training and evaluation; thirdly, the Discussion section highlights implications of the Results and potential future work; this is followed by the Conclusion.

5.2. Method

This section sets out the methods used in this chapter, describing the dataset, the exploratory analysis process and finally the experiments run with machine learning classification models. A diagram of the analysis process can be seen in Figure 6.

5.2.1. The dataset

The previous chapter discussed the process of creating a labelled dataset of text samples taken from Reddit which described posters' experiences of different types of psychological abuse. An annotation scheme was created consisting of six different categories of psychologically abusive behaviour: 1.*Rules, Control and Micro-regulation*; 2. *Justify, Minimize and Deny* abuse; 3. *Threats, Intimidation and Punishment*; 4. *Shaming, Degrading and Ignoring*; 5. *Isolation*; 6. *Surveillance, Monitoring and Harassment*. A more detailed description of each category including examples can be found in the preceding chapter. For brevity, in this chapter the labels will be referred to in the following shorthand: 1. *Rules*; 2. *Justify*; 3. *Threat*; 4. *Shame*; 5. *Isolate*; 6. *Surveillance*.

The full dataset consisted of 28,630 posts, consisting of all posts from three subreddits (r/abusiverelationships, r/domesticviolence, r/emotionalabuse) during 2021, with duplicates, empty or deleted posts discarded, and keeping only the first post from each user. A random sample of 2000 posts was then labelled by four annotators, with all four annotators labelling the first 500 posts, and the remaining 1500 posts being labelled by different combinations of two annotators. The dataset retained the individual labels from each annotator, as well as created two aggregated labels: Aggregated (Wide) included labels which any one annotator had chosen; whilst Aggregated (Narrow) only included labels where at least two annotators had agreed on that label.

As described in the previous chapter, one of the annotators had a low level of inter-annotator agreement with the others, and appeared to be an outlier. For this reason, posts labelled by this annotator were excluded for the purpose of the analysis in this chapter, leaving a dataset of 1500 posts.

The characteristics of the dataset, including the distribution of labels, have been described in the previous chapter, but two aspects deserve highlighting here. Firstly, the dataset is relatively small for a machine learning dataset: the average size of dataset for supervised machine learning problems in the systematic review from Chapter 3 was 73,847 instances. The small size of the dataset requires using specific machine learning techniques, because some models will not learn well from such a small amount of data (Ge et al., 2023). Since machine learning models rely on statistical techniques to pick up on patterns within data, having a large enough dataset to discern key patterns is very important to their performance. If the dataset is too small, they may not have enough information to identify true trends in the data (Wang et al., 2020). As discussed in section 2.3.4 of the Background chapter of this thesis, pre-trained models (such as BERT or DistilBERT) have historically performed well on problems with small amounts of data, also known as "few-shot" learning problems. This is because they encode existing knowledge from related domains to apply it to a new problem, in a process known as Transfer Learning (Wang et al., 2020).

Secondly, the dataset has multiple labels which are highly imbalanced, meaning that the number of positive instances in the dataset is a lot smaller than the number of negative instances. Table 11 shows the absolute number and percentages of positive instances for each label with the Aggregated (Narrow) and Aggregated (Wide) datasets. The most balanced class is the *Threat* class, where 46.6% of texts are positive for this behaviour in the Aggregated (Wide) dataset. *Surveillance, Isolation* and *Rules* are also highly imbalanced, particularly in the Aggregated (Narrow) dataset.

The importance of balanced classes was introduced in Section 2.3.2 of the Background chapter, but to reiterate, it is harder for machine learning models to learn from imbalanced datasets because they tend to learn the characteristics of the dominant class over those of the less dominant class. For example, if a dataset has 50 examples, with 48 of them belonging to the positive class and 2 to the negative class, a machine learning model that simply classified every example it saw with a positive label would already achieve 96% accuracy. This means that different evaluation techniques need to be applied when learning from imbalanced datasets, so that the evaluation measures capture the true performance on the less balanced class. Of particular relevance is the F1 score metric, introduced in Section 3.4.3.2 of the Systematic Review Chapter, which calculates a weighted average between Precision (also known as specificity, or true negative rate) and Recall (also known as sensitivity, or true negative rate), and therefore provides a more accurate assessment of model performance on an imbalanced dataset.

Table 11

Distribution of Labels within the Dataset of Annotators 1, 2 and 4.

	Rules	Justify	Threat	Shame	Isolation	Surveillance
Aggregated (Narrow)	73 (4.8%)	114 (7.6%)	352 (23.4%)	180 (12%)	43 (2.8%)	49 (3.3%)
Aggregated (Wide)	247 (16.4%)	323 (21.5%)	623 (41.5%)	398 (26.5%)	159 (10.6%)	99 (6.6%)

Note: Absolute number of positive instances for each label, with the percentage of the dataset in brackets. The dataset consists of 1500 examples since examples labelled by one annotator were removed due to low inter annotator agreement.

5.2.2. Exploratory Analysis

As introduced in the Background chapter (section 2.3.5), unsupervised machine learning methods, such as clustering algorithms, can be used to analyse a dataset in an exploratory way and to unearth patterns or themes within data that may not be immediately apparent through manual inspection. Unsupervised clustering was applied to the full dataset scraped from Reddit in order to explore whether any linguistic themes emerged organically out of the data.

Unsupervised clustering was conducted using the k-means clustering algorithm on the full dataset of unlabelled Reddit posts (n = 28, 630). This was implemented with the Scikit-learn library using Lloyd's algorithm (Lloyd, 1982) and greedy k-means++ (Arthur & Vassilvitskii, 2007), with bigram TF-IDF vectors as input. The k-means clustering algorithm divides data into k clusters using a mathematical measure of distance within a multi-dimensional abstract space (see Section 2.3.5.1 in the Background chapter for further explanation). Lloyd's algorithm is the most commonly used implementation of k-means which assigns examples to their nearest cluster based on the least squared Euclidean distance and then iteratively updates these assignments until they no longer change (Lloyd, 1982). This approach is simple but not computationally efficient as is not guaranteed to converge (meaning, it is not guaranteed to find the optimum solution). Greedy k-means++ is an adaptation of this algorithm that uses an adaptive sampling approach to selecting candidate clusters. It is more computationally efficient in most cases than Lloyd's algorithm (Bhattacharya et al., 2019).

To prepare the data for input into the k-means algorithm, it was transformed into TF-IDF vectors. These vectors identify the most relevant words in a text, therefore increasing the relevance of the clustering output, because common words that appear in all texts are reduced in importance. A further explanation of TF-IDF vectors can be found in Background chapter section 2.3.6.1. These algorithms were implemented in Scikit-learn, a commonly used Python coding package specifically designed for data science and machine learning applications. The number of clusters was adjusted experimentally between k=6 and k=25, and manually inspected to identify emergence of clear themes.

Additional exploratory analysis was conducted by examining the most frequent uni-, bi-, and tri-grams in each class. A common exploratory analysis technique is to identify the most frequent co-occurring words in a set of documents, also known as collocations. An 'n-gram' refers to a combination of 'n' words appearing together in a set of texts – for example, a combination of two words is referred to as a 'bigram'. This n-gram analysis allowed for an elucidation of what kind of language was being used in each class, which firstly served to sanity check the labelling of each class, and secondly served to demonstrate linguistic themes appearing in the full dataset. The top bi-grams are reported and described in the Results, since they were the most illustrative of the n-gram analyses.

5.2.3. Classification

Since the six labels are not mutually exclusive (each example could be labelled positive for more than one class) the classification problem was framed as six binary classification tasks, one for each label. This means that instead of training one algorithm to recognise whether an example was positive for all six labels at once, six separate algorithms were trained, one for each label.

The Aggregated (Wide) dataset was chosen for training the classification models, since there was a very low number of positive instances in some classes for the Aggregated (Narrow) dataset (see discussion about the importance of balanced classes above). Models for the

Threat label were trained on both the Aggregated (Wide) and Aggregated (Narrow) data in order to compare the outcomes. Furthermore, Models for the Threat and Shame labels were also trained on data from Annotator 1 only, to compare the use of labels from one annotator and to experiment with increasing the amount of data.

Initial experiments were conducted with traditional classifiers using the python Scikit-learn library. The Grid Search Cross Validation function was used to test different parameters and identify the best performing parameters for future tests. Grid search Cross Validation (CV) is a testing algorithm which tries training models with a variety of different model hyper-parameters and then compares the results using CV. Further explanation of CV can be found in the Background Chapter (Section 2.3.7). The results of Grid search CV indicated that 3 different traditional models performed the best and these were selected for testing on each of the labels: Linear SVC, Logistic Regression and Random Forest.

These models were tested using both unigram and bigram TF-IDF vectors as inputs, and subsequently with pre-trained Embeddings from the Spacy English Medium model (*'en_core_web_md'*). TF-IDF vectors and Embeddings are two approaches to vectorisation which are described in the Background Chapter (Section 2.3.6). SpaCy is a widely used Python library that contains pre-trained Embedding models for different languages (SpaCy, 2023). This allows the pre-processing of text to incorporate pre-existing knowledge of language into vectorisation before the text is fed into the machine learning algorithm. Both the TF-IDF and the Embeddings pipelines used pre-processing steps, implemented with the nltk library (Loper & Bird, 2002), which included converting text to lowercase, removing punctuation, URLs and stopwords, and tokenization (where documents are split into separate tokens, which usually means individual words).

Finally, the DistilBERT model from Hugging Face was fine-tuned on data for each task. Hugging Face is a machine learning library that carries implementations of large machine learning models such as BERT and DistilBERT (Wolf et al., 2019) (see Background Chapter Section 2.3.4 for a more detailed introduction to these models). The DistilBERT model was trained using an Adam optimiser (Zhang, 2018) over 3 training epochs. An optimisation algorithm is a component of training a deep learning network that helps the network efficiently converge towards an optimal solution. The number of epochs indicates the number of times that the optimiser runs through all of the training data. More epochs give the model more exposure to the data, which may lead to higher accuracy but increase the risk of overfitting. Whilst an in-depth explanation of Adam lies beyond the scope of this thesis, introductory explanations to machine learning optimisation can be found in Alpaydin (Alpaydin, 2020)

For each experiment, accuracy, precision, recall and F1 score were calculated using a held-out test set, with an 80-20 train-test split for the traditional models, and a 70-15-15 train-test-validation set split for DistilBERT.

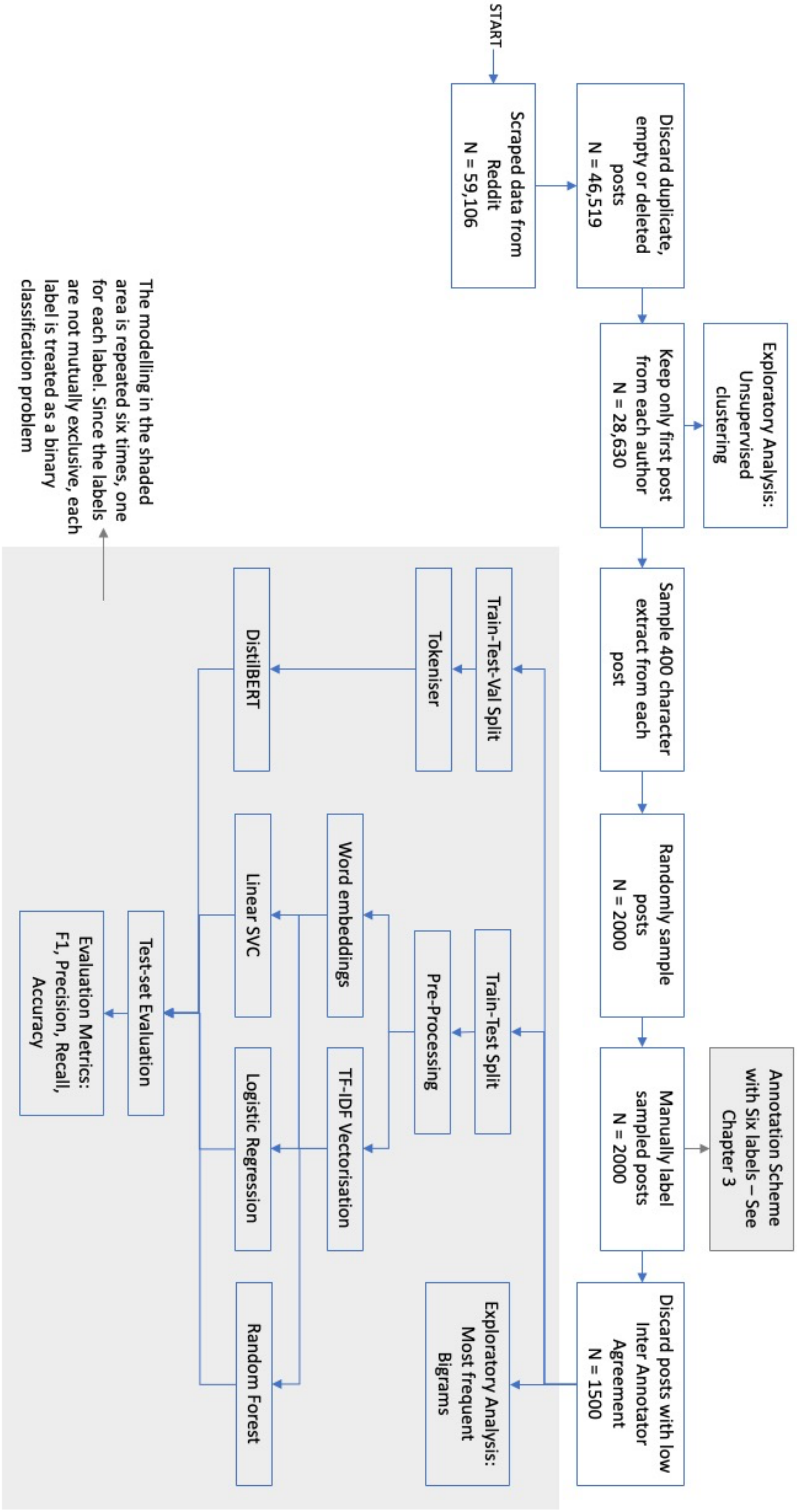

START
Scraped data from Reddit N = 59,106
Discard duplicate, empty or deleted posts N = 46,519
Keep only first post from each author N = 28,630
Exploratory Analysis: Unsupervised clustering
Sample 400 character extract from each post
Randomly sample posts N = 2000
Manually label sampled posts N = 2000
Annotation Scheme with Six labels – See Chapter 3
Discard posts with low Inter Annotator Agreement N = 1500
Exploratory Analysis: Most frequent Bigrams
Train-Test-Val Split
Tokeniser
DistilBERT
Train-Test Split
Pre-Processing
Word embeddings
TF-IDF Vectorisation
Linear SVC
Logistic Regression
Random Forest
Test-set Evaluation
Evaluation Metrics: F1, Precision, Recall, Accuracy
The modelling in the shaded area is repeated six times, one for each label. Since the labels are not mutually exclusive, each label is treated as a binary classification problem

Figure 6 – ***Diagram of Machine Learning Process***

5.3. Results

5.3.1. Exploratory Analysis: Unsupervised Clustering

K-means clustering with TF-IDF bigram vectors was used to identify latent themes within the dataset. The results of unsupervised text clustering were difficult to interpret, although for some k, human-interpretable clusters did emerge. The result of applying K means clustering with n=10 clusters and bigram TF-IDF vectors as features are shown in Table 12 below. Some of the clusters appear to surface relevant human interpretable topics, such as child abuse, remembering past relationships, or break ups.

Table 12

Results of K-means clustering with k=10 clusters

Cluster Theme	Terms
Remembering past relationship	*Long time, don't know, feel like, long time ago, time ago, felt like, don't want, abusive relationship, years ago, didn't want*
Difficult emotions	*Felt like, started crying, like shit, didn't come, trying fix, saying don't, crying saying, half hour, long distance, relationship, don't love*
Effect of past abuse on current relationship	*Throw things, new people, abusive relationship, truly love, physically emotionally, abusive, current relationship, new boyfriend, mental breakdown, hi im, wanted share*
Friend seeking advice	*Abuse like, don't know want, able help, parents house, know want, family friends, need help, don't want, don't know, im constantly*
Uncertainty about abuse	*Don't know, abusive relationship, don't want, emotional abuse, emotionally abusive, years ago, year old, mental health, felt like, best friend*
Break up and no contact	*Told let, don't want, abuse want, got scared, friend got, old friend, broke heart, blocked number, told people, anymore dont*
Domestic violence	*Domestic violence, victim domestic, restraining order, let know, don't know, years ago, abusive relationship, domestic abuse, need help, feel like*
Healing/reconciliation	*Better feel, talk know, trying better, happened told, like felt, apologised said, im lost, im upset, thought id, blah blah*
Emotional abuse	*Feel like, don't know, like im, feel like im, feel like, makes feel, don't want, make feel, abusive relationship, emotionally abusive*
Child abuse / family abuse	*Child abuse, abusive father, end life, dad got, know im, emotionally abusive, don't want, im crying, im currently, im crazy*

5.3.2. Exploratory Analysis: Most common n-grams

Looking at the most common bigrams for the six labels gives insight into which terms are appearing frequently and in which context. Table 13 shows the most common bigrams and their respective counts for each class in the Aggregated (Wide) dataset.

The bigram analysis reveals some interesting insights for each class in the data. The presence of "social, media" in the most common bigrams from the *Surveillance* class chimes with previous research which has indicated that digital technologies, including social media, are frequently used in stalking and harassment behaviour by intimate partners (Freed et al., 2018; Tanczer, López-Neira, & Parkin, 2021). The small size of this class should be noted – there are only 99 posts in the dataset that show *Surveillance* behaviour (see Table 1), so 9 occurrences of "social, media" indicates that it is present in 9.1% of posts in this class. However, other top bi-grams in this class ("don't, know", "feel, like", "even, though") are generic phrases which are present in other classes.

Table 13

Most frequent bigrams in each class

Rules	**n**	**Justify**	**n**	**Threat**	**n**	**Shame**	**n**	**Isolate**	**n**	**Surveillance**	**n**
Even, though	11	Even, though	9	Mental, health	15	Mental, health	11	Don’t, know	15	Don’t, know	13
Don’t, know	11	Years, ago	8	Domestic, violence	12	Best, friend	7	Feel, like	9	Social, media	9
Emotionally, abusive	6	Mental, health	6	Red, flags	11	Calling, names	6	Tried, leave	6	Feel, like	5
Multiple, times	6	Last, night	5	Social, media	11	Anger, issues	5	Family, friends	5	Even, though	4
Makes, feel	5	Nothing, happened	5	Multiple, times	7	Trust, anyone	5	Best, friend	5	Trying, get	4
Ive, tried	5	Bad, guy	4	Depression, anxiety	4	4, years	5	Live, together	4		
3, years	5	Physically, abusive	4	Short, story	4	Came, home	5	Don’t, want	4		
Came, home	5	4, years	4	Panic, attack	4	Emotionally, abused	5	Every, time	4		
One, thing	5	Leave, alone	4	Physically, mentally	4	Months, ago	5				
Every, day	4	Multiple, times	4			Anyone, else	4				
Good, enough	4					Domestic, violence	4				
Last, time	4					Verbal, abuse	4				
Years, old	4					Panic, attack	4				
Say, something	4					Quit, job	4				
Full, time	4					Year, old	4				
Calling, names	4					Silent, treatment	4				
Months, ago	4					Two, days	4				
Spend, time	4					Many, times	4				

In the *Isolate* class, the presence of "family, friends" and "best, friends" in the top bigrams reflects the fact that much isolating behaviour consists of isolating a partner from loved ones. The presence of "Live, together" might indicate that living with a partner often escalates isolating behaviour since the perpetrating partner has more access to, and more control over their victim. Examples of this bigram in this class are: "*she is extremely jealous and constantly accuses me of cheating on her [...] we live together so my home is a hostile environment" (Paraphrased post extract)* and *"he knows that she isn't able to leave so as long as they live together she belongs to him [...] he threatens to take their child if she doesn't have sex with him" (Paraphrased post extract).*

In the Shame class, the presence of "calling, names" and "verbal, abuse" reflects the fact that insults and name-calling are a frequent shaming behaviour used by perpetrators (Duluth Domestic Abuse Intervention Program, 1984). "Anger, issues" may reflect that shaming behaviours from a perpetrator often stem from outbursts of anger and an inability to self-regulate angry feelings (Bancroft, 2003). "Trust, anyone" demonstrates the complex role of trust and the loss of trust common after abusive relationships, illustrated by examples such as: "*I don't trust anyone"*, "*you made me feel I couldn't trust anyone else", "he tells me it's my fault because I don't trust anyone" (paraphrased post extracts)*. "Silent, treatment" is likely to reflect the frequent use of this behaviour by perpetrators as a mechanism of shaming and controlling the victim (Stark, 2009).

Two of the top bigrams in the Justify class exemplify behaviours that are commonly used to justify or deny abuse. Acting like "nothing, happened" reflects an attempt to deny a victim's reality by pretending abuse didn't took place, e.g. "*he calls me names, and now im crying alone in the dark whilst he plays games acting like nothing happened"; "I was scared for my* *life and he acted like nothing happened" (paraphrased post extracts)*. The appearance of "bad, guy" in the top bigrams also reflects the shifting of blame onto the victim*: "in my head, I know he's the bad guy, but he says I need to change a lot"; "he said the all problems we faced are not his fault, and he doesn't feel he is the bad guy here"; "our kids start crying in fear when he starts shouting, but if I take them, I'll be the bad guy for leaving" (paraphrased post extracts)*.

Some of the top bigrams reflect noise in the dataset, which is to be expected in messy, real-world text - for example, the presence of "4 years" and "3 years" in the top bigrams in several classes. A closer look at posts containing "4 years" shows that this is likely to reflect a common phrase pattern when talking about relationships:

> *"this is 4 years after we broke up"*
> *"I faced so much manipulation those 4 years"*
> *"I left an abusive relationship about 4 years ago"*
> *"it's been 4 years of mental and financial abuse at his hands"*
> *"we broke up after 4 years together".*

Additionally, the presence of "years old" and "year old" in the top bigrams seems to be due to posters introducing themselves at the beginning of their post (e.g. "*hi everyone. I'm a 25 year old male*"), and partly due to posters talking about children (e.g. "*we have a three year old together*").

Overall, the presence of "don't know" in the top bigrams of multiple classes does capture the sense of doubt and uncertainty that runs through many posts in the dataset and chimes with qualitative observations made whilst labelling the data. Posts like "*I don't know what to make of his behaviour*" *(Paraphrased post extract)* and "*I don't know who to talk to or where to turn*" *(Paraphrased post extract)* are common in the dataset. Furthermore, the presence of the verb phrase "feel like" also reflects a large number of posts that discuss uncertainty around emotions, such as:

> *"I really feel like she went too far but part of me feels like I deserved it"*
>
> *"He makes me feel like a chore"*
>
> *"He always made me feel like a nuisance"*
>
> *"He told me he feels awful that I'm afraid of him. And that makes me feel like I'm overreacting"*

Overall, examining the top bigrams for each class gives us some insight into common themes and linguistic patterns used by posters when talking about abuse.

5.3.3. Machine Learning: Classifier Results

Table 14

Results of Machine Learning Classifiers by class, using the Aggregated (Wide) labels

Model	**Features**	**Metric**	**Label**					
			1_Rules	**2_Justify**	**3_Threat**	**4_Shame**	**5_Isolate**	**6_Surveil**
		Examples	1500	1500	1500	1500	1500	1500
		Training Examples	1200	1200	1200	1200	1200	1200
		Positive Training Examples	198	258	498	318	127	79
		Positive Testing Examples	49	65	125	80	32	20
RF	**TF-IDF**	Accuracy	0.84	0.77	0.7	0.75	0.87	0.93
		Precision	0.5	0.47	0.67	0.53	0.25	0.5
		Recall	0.2	0.43	0.54	0.46	0.09	0.1
		F1	0.29	0.45	0.6	0.49	0.14	0.17
	Embeddings	Accuracy	0.83	0.79	0.69	0.74	0.89	0.93
		Precision	0	0.67	0.69	0.55	0	0
		Recall	0	0.09	0.49	0.14	0	0
		F1	0	0.16	0.57	0.22	0	0
LR	**TF-IDF**	Accuracy	0.79	0.76	0.7	0.69	0.83	0.94
		Precision	0.38	0.45	0.65	0.44	0.22	0.57
		Recall	0.49	0.55	0.59	0.59	0.25	0.2
		F1	**0.43**	0.5	0.62	0.5	0.24	**0.3**
	Embeddings	Accuracy	0.7	0.69	0.72	0.63	0.63	0.79
		Precision	0.3	0.39	0.65	0.39	0.14	0.13
		Recall	0.61	0.72	0.7	0.69	0.47	0.4
		F1	0.4	0.51	0.67	0.5	0.21	0.2
Linear SVC	**TF-IDF**	Accuracy	0.78	0.73	0.71	0.68	0.82	0.93
		Precision	0.36	0.4	0.68	0.43	0.22	0.5
		Recall	0.49	0.49	0.6	0.65	0.25	0.1
		F1	0.42	0.44	0.64	0.52	0.23	0.17
	Embeddings	Accuracy	0.7	0.66	0.72	0.62	0.63	0.74
		Precision	0.31	0.36	0.66	0.39	0.14	0.12
		Recall	0.69	0.72	0.69	0.72	0.5	0.45
		F1	**0.43**	0.48	0.67	0.51	0.22	0.19
BERT-fine tuning		Accuracy	0.83	0.81	0.83	0.76	0.84	0.92
		Precision	0.41	0.60	0.81	0.63	0.40	0
		Recall	0.39	0.57	0.79	0.51	0.19	0
		F1	0.40	**0.58**	**0.80**	**0.56**	**0.26**	0

Table 14 shows the accuracy, precision, recall and F1-score metrics for the machine learning classifiers by feature type, as well as the number of positive instances in the training and testing datasets for each class.

The six classification tasks presented here all involve identifying if the text in question contained a type of psychologically abusive behaviour. In general, texts containing the abusive behaviour are less common than those not containing the abusive behaviour (apart from in the threat class, where the number of positives and negative instances is almost equal). In general this suggests that false negatives are more likely than false positives, so the expectation is that recall scores would be higher than precision scores. This is the case with the results from the Logistic and Regression and Linear SVC algorithms – however, the results from the Random Forest model reverse this trend. In BERT fine-tuning, precision is slightly higher than recall but in three classes is about the same.

When examining precision and recall, it is important to think about the downstream applications of the task and whether or not an end user of such a model would care more about false positives (where a text is wrongly predicted as containing descriptions of abuse when it does not) or false negatives (where a text contains descriptions of abuse but it is predicted as not containing any mention of abuse). For example, in machine learning for medical imaging, where ML algorithms are used to predict the presence of a tumour on a scan, false negatives are far less desirable than false positives, because of the high consequences of accidentally missing a life-threatening disease.

When it comes to the classification of description of abuse, it is likely that downstream applications would prefer false positives over false negatives, because decisions are likely to be examined by a human who is able to correct any false positives, whereas if false negatives predominate then there is a risk of missing abusive behaviour within a dataset. We would therefore prefer that recall remains high. From this perspective, Logistic Regression and Linear SVC with embedding both perform relatively well, even though their F1 scores are low, because they have high recall scores and low precision scores.

As can be seen from the table, some classes achieved high accuracy but low F1 scores with classifiers (for example, *Isolate* achieved 0.83 accuracy but 0.24 F1 score with logistic regression trained on TD-IDF vectors). Due to the large class imbalance with most of the labels (there are many more examples that don't contain the behaviour than those that do), accuracy is not a very reliable metric for assessing the outcome of the model. In a highly imbalanced dataset, a model that predicted every example as negative would still 'predict' the outcome with high accuracy, simply because of the presence of many negative examples. Therefore, for this data it is more important to look at the precision, recall, and F1 scores.

The F1 scores are very low across most labels, apart from *Threat*. This is reflective of the small number of training examples, the high imbalance of classes, and the relative complexity of the phenomenon to be classified. To improve these scores, it is likely that a significantly larger number of training examples would be needed.

For the *Threat* label, both Logistic Regression and Linear SVC with Embeddings performed reasonably well, achieving 0.67 F1 score. In general, the Logistic Regression and Linear SVC models outperformed the Random Forest models. Embeddings seemed to improve performance slightly over TF-IDF vectors in some cases, but not all.

The DistilBERT model was a significant improvement in the *Threat* category, achieving 0.8 performance. In most other categories it did improve on the performance from the traditional models, but in some cases only by a small amount (e.g. *Shame*: 0.56 (DistilBERT) vs 0.52 (Linear SVC with TF-IDF)).

5.3.3.1. Training with Different Data

In order to understand the impact of the number of training examples on the performance of the models, training was also conducted for the *Threat* and *Shame* labels on a larger dataset (n=2000) of examples using only the annotations from Annotator 1 (A1). Models were also trained on the Threat label task using the Aggregated (Narrow) dataset, to see how this performed in comparison to the Aggregated (Wide) dataset. The results from these additional training runs can be seen in Table 15.

Using the larger dataset (A1) resulted in a slight improvement of F1 score (+0.1) for the Threat class, but a reduction in F1 score (-0.7) for the Shame class. This might be because, even though the A1-only annotated dataset is 500 instances larger, it has very similar numbers of positive instances in the training and test sets for the Shame class.

Furthermore, using the Aggregated (Narrow) dataset for the Threat label resulted in a significant reduction in F1-score for all the classifiers. This reflects why the Aggregated (Wide) dataset was chosen to be the main experimental dataset for the current analysis - the Aggregated (Narrow) dataset has a significantly smaller number of positive instances, which is likely to blame for the decrease in performance.

5.3.3.2. Training with More Data

Figure 7 shows a graph of the number of positive training and testing examples plotted against the F1 score of the DistilBERT fine-tuned model. The data points represent the six labels from the Aggregated (Wide) dataset shown in Table 14, and the two A1-only datasets shown in Table 15, which are circled on the graph. The trendlines indicate that the number of positive training and testing examples is directly correlated with the performance of the best performing classifier. This suggests that increasing the amount of labelled data is likely to increase the performance of the model even for currently poorly performing labels.

However, the circled points, representing data labelled by only one annotator, fall slightly below the trendline. The A1-only dataset contained 500 more instances than the Aggregated datasets, but only contains the labels from a single annotator. This hypothetically could have meant a more consistent definition for the label, which could have lead to a better performing model. However, as can be seen from the circled points in the figure, the A1-only dataset offered only very incremental increases in performance as compared to the smaller dataset with multiple annotators. This suggests that having multiple annotators enriches the data and makes it more robust, and that a dataset with a single annotator is not as effective as a dataset with aggregated labels from three annotators.

The DistilBERT model offered reasonable performance for the label with the largest number of positive training instances (*Threat*, n=498). This suggests that fine-tuning is a good approach to use for this type of classification problem, and could be applied to the other labels if they had more positive instances. If we extrapolate that approximately 500 positive instances are required to achieve reasonable performance with DistilBERT, and use the percentages of positive instances in the current dataset, then achieving similar performance for the *Surveillance* label would require labelling approximately 7,600 instances. Given the labelling time for the current dataset was between 50-100 instances per hour (excluding time taken for annotator training), this is likely to take approx. 76 to 152 hours of work for future annotators. This demonstrates the difficulties with recruiting expert annotators, who also need to be trained, to do this type of fine-grained, complex annotation work.

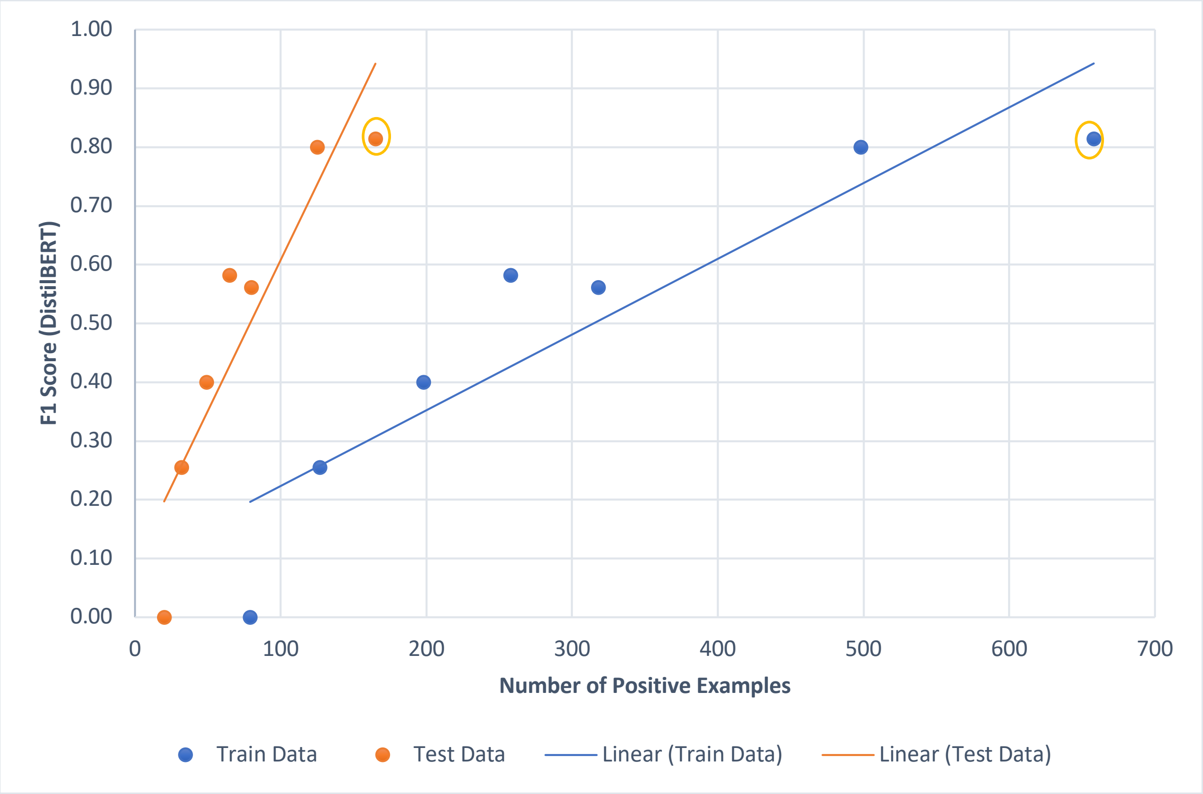


Figure 7 – ***Number of Positive Training and Testing Examples and F1-score of DistilBERT fine-tuned, by label***

Table 15

Results from Training Classifiers with Different Datasets

Model	Features	Metric	Threat (Wide)	Threat (A1)	Threat (Narrow)	Shame (Wide)	Shame (A1)
		Examples	1500	2000	1500	1500	2000
		Training Examples	1200	1600	1200	1200	1600
		Positive Training Examples	498	658	282	318	326
		Positive Testing Examples	125	165	70	80	81
		Level of agreement	0.61	N/A	0.61	0.44	N/A
RF	TF-IDF	Accuracy	0.7	0.73	0.79	0.75	0.79
		Precision	0.67	0.71	0.71	0.53	0.44
		Recall	0.54	0.56	0.46	0.46	0.23
		F1	0.6	0.63	0.56	0.49	0.31
	Embeddings	Accuracy	0.69	0.7	0.79	0.74	0.8
		Precision	0.69	0.67	0.67	0.55	0.43
		Recall	0.49	0.57	0.23	0.14	0.04
		F1	0.57	0.61	0.34	0.22	0.07
LR	TF-IDF	Accuracy	0.7	0.73	0.83	0.69	0.69
		Precision	0.65	0.67	0.65	0.44	0.27
		Recall	0.59	0.68	0.57	0.59	0.31
		F1	0.62	0.68	0.61	0.5	0.29
	Embeddings	Accuracy	0.72	0.72	0.75	0.63	0.66
		Precision	0.65	0.65	0.47	0.39	0.34
		Recall	0.7	0.72	0.67	0.69	0.69
		F1	0.67	0.68	0.55	0.5	0.45
Linear SVC	TF-IDF	Accuracy	0.71	0.72	0.81	0.68	0.66
		Precision	0.68	0.66	0.59	0.43	0.26
		Recall	0.6	0.67	0.59	0.65	0.37
		F1	0.64	0.66	0.59	0.52	0.31
	Embeddings	Accuracy	0.72	0.72	0.74	0.62	0.64
		Precision	0.66	0.64	0.45	0.39	0.32
		Recall	0.69	0.72	0.64	0.72	0.7
		F1	0.67	0.68	0.53	0.51	0.44
BERT-fine tuning		Accuracy	0.83	0.84	0.87	0.76	0.81
		Precision	0.81	0.87	0.56	0.63	0.60
		Recall	0.79	0.76	0.92	0.51	0.42
		F1	**0.80**	**0.81**	**0.69**	**0.56**	**0.49**

5.4. Discussion

The results presented above demonstrate that the use of computational text mining tools can be useful to explore and understand how DA appears in large datasets sourced from social media platforms like Reddit. This section addresses the research questions presented in the Introduction by discussing the following aspects: first, the insights which arose from the exploratory analysis; second, the most successful models at classifying types of psychological abuse; and third, which types of psychological abuse were classified most successfully. This is followed by a discussion of limitations and potential future work.

5.4.1. Exploratory Analysis

Exploratory analysis using k-means clustering and n-gram analysis provided insight into the ways users talked about psychological abuse on Reddit.

Some of the k-means clusters reflected topics which had been anecdotally observed during labelling, such as how the experience of abuse in a past relationship can continue to have a negative effect on a new relationship even after the abuse has ended ("new people, abusive relationship, truly love, physically emotionally, abusive, current relationship, new boyfriend, mental breakdown"…). Another topic seemed to reflect some of the heartbreakingly difficult emotions that arise during the ups and downs of abusive relationships ("Felt like, started crying, like shit, didn't come, trying fix, saying don't, crying saying, half hour, long distance, relationship, don't love"). Another topic possibly reflected the 'cycle of abuse' (Pandora Project, 2023), where apologies and reconciliation is often followed by further abuse, that is typical of many abusive relationships ("Better feel, talk know, trying better, happened told, like felt, apologised said, im lost, im upset, thought id, blah blah"). Whilst k-means clustering did seem to offer some initial insight into latent topics in the dataset, choosing the correct k was achieved through manual trial-and-error. Future work could use unsupervised evaluation methods such as Rate of Perplexity Change (RPC) (Xue et al., 2019), or combine k-means clustering with another topic modelling method such as Latent Dirichlet Allocation (LDA) in order to make these results more robust.

N-gram analysis gave a good indication of the kinds of language present in each label and across labels. The presence of "feel like" and "don't know" across labels seemed to reflect the uncertain and emotional nature of experiencing DA. The presence of "social media" in the top bi-grams of the *Surveillance* class reflected the use of technology by perpetrators to watch and harass victims (Freed et al., 2018). The presence of "bad guy" and "nothing happened" in the top bi-grams for the *Justify* class indicated the shifting of blame and denial of the victim's reality that is common in psychologically abusive relationships (Stark, 2009). Overall, n-gram analysis offered useful insights such as these, and provided a sanity check as to the robustness of the labels, since the top bi-grams generally reflected the underlying concept of each label.

5.4.2. Most successful models

DistilBERT showed the best performance across all the models in all but two of the labels. Due to its track record with few-shot classification problems, and the complexity of the task, it is not surprising that DistilBERT out-performed traditional models. These results reflect those of Botelle et al., who found that BioBERT, an adaptation of BERT for medical text classification, outperformed other models at detecting different types of abuse in electonric health records (Botelle et al., 2022).

5.4.3. Types of psychological abuse

The type of abuse which ML models classified most successfully was *Threatening, Intimidating and Punishing* behaviour, which consistently achieved the best F1 score for all labels across all the different types of models which were tested. This is most likely due to this label being the most balanced of all abuse types – 41.5% of examples were positive for Threat in the Aggregated (Wide) dataset. Since *Threatening, Intimidating and Punishing* behaviour is the most common type of psychological abuse reported on Reddit, it is promising that classifiers perform well at detecting it. A classifier that could detect reports of threatening behaviour could be useful for future research and for practitioners seeking to understanding the prevalence of reports of threatening behaviour in their own existing data.

Threatening behaviour is also most likely the closest, linguistically speaking, to other forms of hate speech and abuse which have already been the subject of extensive research in computational linguistics. For example, there is a significant body of research and practice which has designed machine learning classifiers to detect hate speech and online abuse, often motivated by tech companies' desire to moderate harmful content on their platforms (Mishra, Yannakoudakis, & Shutova, 2019). Since the BERT model, which DistilBERT is based on, has previously been used successfully to detect abusive speech online (P. Liu, Li, & Zou, 2019; Mozafari, Farahbakhsh, & Crespi, 2020), it is possible that it already has some 'understanding' of abusive content. This kind of speech is likely to be closer to the *Threat* label than it is to the other labels in the dataset, which capture more nuanced kinds of abuse that are specific to DA.

The low performance on the *Rules* label from all classifiers label mirrors the fact that it was a tricky label to annotate, and achieved the lowest IAA out of all labels. This is likely to be because, whilst controlling and micro-regulating behaviour is common, descriptions of controlling behaviour tend to be implicit and contextual. For example:

> *"A friend called me asking me to go to a New Year's party. But she snatched the phone away from me and told my friend: "He can't go. End of conversation." And hung up."*
>
> *"It's so sad to watch my strong sister changing her own behaviour in case it causes him to explode. She's gone as far as changing where the dog sleeps because he told her to."*

These examples both describe controlling behaviour according to our annotation guidelines: both describe conduct which causes the victim-survivor to change their behaviour, and indicate 'issuing orders and demanding obedience', which was one of our example behaviours for the *Threat* label. However, there are few linguistic cues of control in these examples for an ML classifier to pick up on.

This contrasts with more explicit examples of *Rules*, such as:

> "*He never lets me wear what I want to, I have to be covered from head to toe."*
>
> *"He doesn't let me go to dance classes or the gym anymore."*
>
> *"He prevents them from leaving the house without his permission. He controls their money and makes them beg for cash for essentials."*

These examples of controlling behaviour contain clear verb patterns like "*doesn't let", "never lets", "controls", "prevents"*, which, hypothetically, might make them easier for an ML model to pick up on. However, it is difficult to confirm this hypothesis since the decision-making mechanism of a deep transformer model like BERT is extremely complicated and opaque, so it is not possible to verify what linguistic features it was using in its decision-making function. This is one of the downsides of using deep learning models, which, whilst powerful, are difficult to explain in a human-interpretable manner.

In contrast to the *Rules* label, many examples of *Surveillance* in the dataset are actually quite explicit (*"He has all my passwords", "He still periodically harasses me via text", "He called my phone 20 times", "My husband has cameras everywhere and a tracker on my car", "he was always checking my phone"*) and contain verbs and nouns that might be common to many examples of Surveillance *("password", "text", "harasses", "called", "cameras", "tracker", "checking"*). Perhaps for this reason, *Surveillance* was relatively easy for human annotators to recognise, and achieved the second highest level of IAA. It was therefore slightly surprising that *Surveillance* achieved such low F1 scores, but this is likely to be due to the very low number of positive examples for this label.

5.4.4. Limitations

Overall, the classifiers presented here were not particularly successful at identifying psychological abuse. This is likely to be because of the complex and nuanced nature of psychological abuse, about which not even human annotators can agree, as discussed at length in the previous chapter. Furthermore, there was not enough labelled data to achieve good performance of the models, and the analysis presented here has indicated that increasing the amount of labelled data could significantly increase performance.

The classifiers presented here do not currently perform well enough to be considered appropriate for the downstream tasks - such as research, policing or advocacy – suggested at the beginning of this chapter. However, they achieve a baseline level of functionality, especially with the more balanced labels, which could be built upon in future work. In particular, the DistilBERT fine-tuned model achieved good performance (F1=0.81) for the

Threat label. Whilst it did not perform as well at identifying other types of psychological abuse, a classifier that can automatically identify descriptions of threatening behaviour could be useful for identifying individuals at risk of this kind of abuse.

5.4.5. Future Work

Besides future work that uses the existing annotation scheme to label more data, a number of additional techniques could be tried that are beyond the scope of this chapter. Data augmentation techniques are popular for increasing the amount of training data (Wei & Zou, 2019) and could be used to increase model accuracy. Custom deep learning models, such as LSTM or GRU models, could also be experimented with, since these have provided state-of-the-art performance in similar text classification tasks (S. Subramani et al., 2019). An LSTM pipeline was experimented with for this chapter, but due to the small amount of data it was difficult to achieve convergence with this model.

Finally, with recent leaps forward in the capabilities of large language models (LLMs) such as GPT-3 and GPT-4 (OpenAI, 2023), it is likely that text classification research will undergo transformation in the coming years. GPT-4 is already being used for data augmentation with small datasets (Møller, Dalsgaard, Pera, & Aiello, 2023) like the one presented in this thesis. Emerging techniques in natural language understanding with LLMs could build on the research presented in this chapter, and may lead to models that can automatically classify different types of psychological abuse in future.

5.5. Conclusion

This chapter examined the application of machine learning models to the dataset of psychologically abusive behaviours presented in the last chapter. Three types of traditional machine learning model (Linear SVC, Logistic Regression and Random Forest) were trained, and a transformer-based model (DistilBERT) was fine-tuned, on six different binary classification tasks according to the six different binary labels of the psychological abuse dataset.

The models achieved good performance, measured by F1 score, on the *Threats, Intimidation and* Punishment label (0.81) but the other labels had too few positive examples to achieve good outcomes. Future work could consider labelling more data or experimenting with data augmentation approaches, as well as the use of other types of deep learning models.

Conclusion

This thesis sought to contribute to the understanding of DA by using computational methods to study reports of psychological abuse in a large dataset of public social media posts. This final chapter summarises the main findings and contributions of the thesis, as well as its limitations, its implications for policy and practice, and suggested future work.

6.1. Main Findings

The research questions presented in Chapter 1 were as follows:

> RQ1: How has existing work has used computational text analysis methods to research domestic abuse?
>
> RQ2: How can we build on existing research to create a dataset of reported psychological abuse in online forums?
>
> RQ3: How successfully do machine learning models learn to classify psychological abuse?

6.1.1. Systematic Literature Review

Chapter 3 responded to RQ1 through a systematic review of existing literature which used computational text analysis methods to study DA. The survey identified an emerging body of work, consisting of 22 unique studies, which used computational methods and looked at some aspect of DA. The included studies drew data from a wide range of sources including social media, police reports, court documents and narratives collected directly from victim survivors. The volume of research in this inter-disciplinary area appeared to be increasing in recent years. The survey looked at the different techniques and study designs used by existing research. Deep Learning and Transformer based models, such as CNNs, RNNs and BERT, appeared to achieve good results in a number of studies; however, their drawback is their high level of opacity making their decision-making mechanisms difficult to understand, which may be undesirable in sensitive environments such as those dealing with DA.

The review also developed a unique 21-item questionnaire to evaluate interdisciplinary machine learning research about DA, because existing quality metrics were difficult to apply to research with both machine learning and DA elements. The 21-item checklist includes points addressing both the quality of the computational aspects of research (e.g. appropriate model used for hypothesis, different models tested and compared) and the DA aspects of research (e.g. definition of violence discussed, clearly described and motivated DA-related hypothesis). This checklist built on existing frameworks for assessing the quality of both ML and mixed methods research (Dreisbach et al., 2019; Hinds et al., 2021; Hong et al., 2018; Siebert et al., 2020; Zhai et al., 2020) and represents a contribution that future researchers could use as a starting point to design their own interdisciplinary studies.

The systematic review ultimately identified a gap in the literature – whilst some of the studies included psychological abuse as a type of DA in their data, none of them sought to classify psychological abuse into different types of psychologically abusive behaviour. This is despite research suggesting that psychological abuse is the most common type of DA (Home Office, 2021). This gap therefore motivated the research described in Chapter 4.

6.1.2. Explainable Dataset of Psychologically Abusive Behaviours

Whilst existing literature has explored how to create machine learning datasets in a transparent and ethical way (Kapania et al., 2023; Muller et al., 2021; Röttger et al., 2021), very limited research exists which actually *creates* explainable datasets. Chapter 4 explained the process of creating a dataset of Reddit posts using an innovative *data explainability* process with three stages. The first stage created an annotation scheme from a literature review of existing research about psychological abuse. By grounding the labels in existing research, the study attempted to make the process of conceptualising the annotations more transparent and robust. This resulted in a pilot annotation scheme that was then refined through expert discussion, as in the process conducted by, among others, Botelle et al. (Botelle et al., 2022). The final stage consisted of measuring and reporting disagreement amongst annotators in the final dataset, to make clear to downstream users of the data where disagreements and ambiguities occurred, as suggested by Röttger (2021).

This process resulted in the creation of a six-label annotation scheme, which represents a new conceptualisation of types of psychological abuse building on existing work (Duluth Domestic Abuse Intervention Program, 1984; Marshall, 1996; Stark, 2009). A dataset of posts from Reddit (n=2000) was then labelled using the annotation scheme by a team of four annotators.

Ultimately, the process presented in this Chapter aimed to answer RQ2 and resulted in a dataset of psychologically abusive behaviours as reported by Reddit forum users.

6.1.3. Using Machine Learning Classifiers to Identify Different Types of Psychologically Abusive Behaviour

Chapter 5 explored the application of machine learning methods to automatically classify the different types of psychological abuse as defined in Chapter 3, thereby answering RQ3 above. Overall, machine learning models were not successful at classifying most types of abuse, because the annotation scheme was too granular for the amount of labelled data, meaning that some classes were highly imbalanced. However, the DistilBERT pre-trained model did perform well (F1-score = 0.81) at classifying one type of psychological abuse: *Threatening, Intimidating and Punishing* behaviour. Furthermore, given that downstream use cases of such a model are likely to prioritise minimising false negatives, the Logistic Regression and Linear SVC models also performed well, since they had high recall rates despite lower F1 scores.

Analysis presented in this chapter demonstrated that the number of positive training and testing examples was strongly correlated with model performance, indicating that increasing the amount of labelled data would likely lead to better performance with classifiers for other types of psychological abuse. This correlates with research in other domains which suggests that machine learning models find it difficult to perform well on datasets of such small size (Wang et al., 2020). In future, the further development of Large Language Models (LLMs) like GPT-4 is likely to quickly revolutionise this field and make few-shot learning significantly more viable. This Chapter therefore provided a starting point for future researchers to apply such emerging techniques to detect psychological abuse.

Overall, the thesis successfully answered the research questions presented in Chapter 1, but finds that machine learning classifiers need more data, and deeper conceptual understanding of the definitions of different types of abuse, to successfully classify different types of psychological abuse.

6.2. Limitations

The main limitation of the thesis as a whole is that its conclusions are limited by the small size of the dataset described in Chapter 3. A larger dataset would likely have enabled better inter annotator agreement between annotators and better model performance. The creation of a larger dataset was mainly hampered by time constraints. To create the 2000-instance dataset presented in this thesis, three researchers spent between 10 and 15 hours labelling data, as well as attending three 2 hour workshops, conducting approximately 3 hours of labelling during training, and multiple check-in meetings during the labelling process. The author spent approximately 35 hours labelling the whole dataset. The average labelling speed was around 50 instances per hour after initial training. Skilled researcher time is difficult to come by and the sensitive and sometimes distressing nature of the data meant that time spent labelling needed to be minimized and spread out so as to reduce the possibility of harm for the researchers. Future work could recruit a larger annotation team or procure a budget to pay researchers for additional time spent labelling data.

The limitations of each chapter are now discussed individually.

Chapter 2, the Systematic Literature Review, was limited by its search strategy, which didn't capture grey literature – such as reports commissioned by DA charities – and some academic literature, such as book chapters, which are not indexed in academic databases. Furthermore, the Quality Assessment criteria used in the review were developed by two of the co-authors (Lilly Neubauer and Isabel Straw) and have not been independently evaluated on other research, so any conclusions drawn about the quality of included studies may be subject to bias.

Chapter 3, the creation of the Explainable Dataset of Psychological Abusive Behaviour, was limited by a lack of diversity in annotators, a lack of insight into the decisions of some annotators, and a low level of agreement between annotators despite discussions over multiple discussions. The number of labelled examples is also probably too small to achieve high levels of agreement on some classes which were highly imbalanced.

Chapter 4 had some technical limitations. Ultimately, the classifiers presented in Chapter 4 don't perform well enough to be useful in downstream tasks. This is likely to be because of the small size and imbalanced nature of the dataset. However, due to time constraints, several techniques that might have improved the performance of the classifiers were not able to be explored: for example, data augmentation, or further few-shot deep learning techniques. Finally, the k-means clustering presented in this paper represented a relatively shallow exploratory analysis – with more time, other algorithms such as LDA could have been explored, and the outcome of k-means clustering could have been quantitatively evaluated to determine the optimum number of clusters.

6.3. Contributions to the Literature

The contributions of this thesis to the literature are as follows:

1. The Systematic Literature Review described in Chapter 2 is the first of its kind to examine computational text analysis methods in intimate partner violence research. This provides a foundation for researchers wanting to use similar methods, and contributes to the development of the field by encouraging and facilitating the use of new, innovative research methods.
2. The Annotation Scheme of six types of psychologically abusive behaviour, presented in Chapter 3, represents a contribution to research about psychological abuse. The annotation scheme built on existing measures and frameworks of psychological abuse to develop a new typology and further extend this into a detailed annotation scheme for labelling machine learning datasets. This is, to the knowledge of the author, the

first annotation scheme for types of psychologically abusive behaviours that has been presented in the literature. This six-label categorisation could be used to label further datasets and built on by researchers wishing to conceptualise psychological abuse for their own projects.

3. The labelled dataset presented in Chapter 3 represents a contribution as the first machine learning dataset of psychologically abusive behaviours.
4. The machine learning analysis in Chapter 4 contributes to the literature by being the only current study, to the awareness of the author, to attempt to use machine learning to classify different types of psychological abuse.

6.4. Implications for Research and Practice

Despite the limitations mentioned above, this thesis does offer implications for future research and practice in this domain.

6.4.1. Implications for Practice

As discussed in the Introduction and Background sections of this thesis, DA is widespread and has an enormous negative impact on victims. Some types of DA are criminal offences in the UK and other parts of the world, including physical violence, sexual assault, stalking, psychological abuse (which is criminalized under Coercive Control legislation in the UK) and more recently, revenge porn (which was made a criminal offence in the 2023 Online Safety Bill). Many organizations within the domestic abuse advocacy sector, as well as police forces and government bodies, are working tirelessly to try and reduce the occurrence, impact and reoccurrence of DA.

However, it is very difficult to effectively reduce a crime when it is not even well understood how often and in what ways it is occurring. Particularly when it comes to psychological abuse, the research still presents more questions than answers – how common is psychological abuse, given that much of it is not reported to the police or other authorities? What kinds of psychologically abusive techniques are abusers using? How do victims of psychological abuse conceive of and talk about their experiences? The knowledge generated in this thesis contributes one brick to building a wall of understanding of psychological abuse, by providing tools and insight into how victim-survivors of psychological abuse talk about their experiences in an online forum.

Firstly, a version of the machine learning models presented in this thesis could be used by law enforcement agencies or researchers to understand what types of psychological abuse are reported in other kinds of dataset (such as in databases of victim statements held by the police or advocacy organisations, for example). This could help such agencies to understand how to target educational interventions or public awareness campaigns aimed at raising awareness of different types of psychologically abusive behaviour.

Secondly, the six-type conceptualisation of types of psychological abuse presented in this thesis could be used by organisations to help train staff members to recognise and categorise different types of psychological abuse. Psychological abuse is not well understood by the general public and even by agencies working with victim-survivors of abuse (B. Barocas et al., 2016; Follingstad, 2007), so the results of the work presented in this thesis provide a clear and well explained reference point for people to learn about how psychological abuse presents itself.

6.4.2. Implications for Research

Many social scientists, including those studying DA, are keen to exploit modern computational methods to make their work more efficient and open up new sources of data. This thesis offers a starting point for DA scholars to use similar methods in their own work, and has deliberately phrased descriptions of computational methods to make them more accessible to a non-technical audience. Care has also been taken at each stage not to oversell computational methods, and to highlight the potential pitfalls and ethical challenges of using advanced algorithms in research.

For DA advocacy and support sector organisations, the findings of the thesis have implications as follows: Chapter 2 offers a jumping off point for organisations wishing to understand the potential for computational methods and how these might facilitate research with the data that they already hold. Furthermore, findings in Chapter 3 help illuminate common types of psychologically abusive behaviour, which could be useful for practitioners working with victim-survivors of psychological abuse. Overall, the thesis has illustrated how existing data can be leveraged, using computational methods, to provide new insights about DA that can help to inform practice in this field.

For computer scientists and data scientists, the thesis offers insight into using computational methods in a social science context. Whilst the thesis has used existing algorithms in its analysis, its innovative aspect comes from the application of existing techniques to a new domain. Particularly, Chapter 3 offer lessons and best practice for labelling new machine learning datasets, especially in topics where humans find it difficult to agree, and demonstrates that conceptual agreements about underlying topics need to be settled before data can be labelled effectively.

Finally, the limitations of each chapter have demonstrated that the state of debate and understanding within the DA field currently limits the efficacy of computational solutions in DA research and practice. This emerging interdisciplinary area offers potential for further research, but more data and discussion is needed before computational tools can be effectively used to fully detect and understand DA in real-world applications.

References

Adily, A., Karystianis, G., & Butler, T. (2021). Text mining police narratives to identify types of abuse and victim injuries in family and domestic violence events. *Trends and Issues in Crime and Criminal Justice* (630), 1-12.

Akbani, R., Kwek, S., & Japkowicz, N. (2004). *Applying support vector machines to imbalanced datasets.* Paper presented at the European conference on machine learning.

Allen, K., Davis, A. L., & Krishnamurti, T. (2021). Indirect Identification of Perinatal Psychosocial Risks from Natural Language. *IEEE Transactions on Affective Computing*.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine learning, 75*, 245-248.

Alpaydin, E. (2020). *Introduction to machine learning*: MIT press.

Alvarez, R. M. (2016). *Computational social science*: Cambridge University Press.

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert systems with applications, 88*, 402-418.

Annapragada, A. V., Donaruma-Kwoh, M. M., Annapragada, A. V., & Starosolski, Z. A. (2021). A natural language processing and deep learning approach to identify child abuse from pediatric electronic medical records. *PLoS One, 16*(2), e0247404.

Arango, A., Pérez, J., & Poblete, B. (2019). *Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation*. Paper presented at the Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France. https://doi.org/10.1145/3331184.3331262

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine, 36*(1), 15-24.

Arthur, D., & Vassilvitskii, S. (2007). *K-means++ the advantages of careful seeding.* Paper presented at the Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.

Australian Bureau of Statistics. (2013). Bridging the data gaps for family, domestic and sexual violence. Retrieved from https://www.abs.gov.au/statistics/people/crime-and-justice/bridging-data-gaps-family-domestic-and-sexual-violence/latest-release#improving-the-evidence-base. Accessed July 2022.

Avieli, H. (2021). False Allegations of Domestic Violence: A Qualitative Analysis of Ex-Partners' Narratives. *Journal of family violence*, 1-13.

Babvey, P., Capela, F., Cappa, C., Lipizzi, C., Petrowski, N., & Ramirez-Marquez, J. (2021). Using social media data for assessing children's exposure to violence during the COVID-19 pandemic. *Child Abuse Negl, 116*(Pt 2), 104747.

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures, 16*(1), 1-18. doi:10.1080/19312458.2021.2015574

Bancroft, L. (2003). *Why does he do that?: Inside the minds of angry and controlling men*: Penguin.

Barocas, B., Emery, D., & Mills, L. G. (2016). Changing the domestic violence narrative: Aligning definitions and standards. *Journal of family violence, 31*(8), 941-947.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California law review*, 671-732.

Basile, K. C., Black, M. C., Breiding, M. J., Chen, J., Merrick, M. T., Smith, S. G., . . . Walters, M. L. (2011). *National intimate partner and sexual violence survey: 2010 summary report*. Retrieved from

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). *The pushshift reddit dataset.* Paper presented at the Proceedings of the international AAAI conference on web and social media.

Bechmann, A., & Zevenbergen, B. (2019). AI and Machine Learning: Internet Research Ethics Guidelines (IRE 3.0 6.1). *Association of Internet Researchers Internet Research: Ethical Guidelines 3.0*, 33.

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater Reliability in Systematic Review Methodology: Exploring Variation in Coder Decision-Making. *Sociological Methods & Research, 50*(2), 837-865. doi:10.1177/0049124118799372

Bhattacharya, A., Eube, J., Röglin, H., & Schmidt, M. (2019). Noisy, greedy and not so greedy k-means++. *arXiv preprint arXiv:1912.00653*.

Blackwell, L., Dimond, J., Schoenebeck, S., & Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction, 1*(CSCW), 1-19.

Botelle, R., Bhavsar, V., Kadra-Scalzo, G., Mascio, A., Williams, M. V., Roberts, A., . . . Stewart, R. (2022). Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. *BMJ Open, 12*(2), e052911. doi:10.1136/bmjopen-2021-052911

Campbell, D. W., Campbell, J., King, C., Parker, B., & Ryan, J. (1994). The reliability and factor structure of the Index of Spouse Abuse with African-American women. *Violence and Victims, 9*(3), 259-274.

Campbell, J. C., Webster, D. W., & Glass, N. (2009). The danger assessment: Validation of a lethality risk assessment instrument for intimate partner femicide. *JOURNAL OF INTERPERSONAL VIOLENCE, 24*(4), 653-674.

Casquilho-Martins, I., Belchior-Rocha, H., & Moro, S. (2022). Unfolding Social Work Research to Address the COVID-19 Impact: A Text Mining literature Analysis. *The British Journal of Social Work, 52*(7), 4358-4377.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). *Mean birds: Detecting aggression and bullying on twitter.* Paper presented at the Proceedings of the 2017 ACM on web science conference.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chu, T., Su, Y., Kong, H., Shi, J., & Wang, X. (2021). Online Social Support for Intimate Partner Violence Victims in China: Quantitative and Automatic Content Analysis. *VIOLENCE AGAINST WOMEN, 27*(3), 339-358.

Crootof, R. (2019). "Cyborg Justice" and the Risk of Technological-Legal Lock In. *Columbia Law Review, 119*(7), 233-251.

Crown Prosecution Service. (2017). Legal Guidance on Controlling or Coercive Behaviour in an Intimate or Family Relationship. Retrieved from https://www.cps.gov.uk/legal-guidance/controlling-or-coercive-behaviour-intimate-or-family-relationship. Accessed 30th May 2022.

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics, 10*, 92-110.

DeepLearning AI. (2021). A Chat with Andrew on MLOps: From Model-centric to Data-centric AI. Retrieved from https://www.youtube.com/live/06-AZXmwHjo?feature=share. Accessed 27th July 2023.

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

Dignum, V. (2017). Responsible artificial intelligence: designing AI for human values. *International Telecommunication Union Journal: ICT Discoveries, Special Issue, 1*, 1-8. doi:https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P01-PDF-E.pdf

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2). doi:https://doi.org/10.1177/2053951715602908

Dobash, R. P., Dobash, R. E., Wilson, M., & Daly, M. (1992). The myth of sexual symmetry in marital violence. *Social problems, 39*(1), 71-91.

Dokkedahl, S., Kok, R. N., Murphy, S., Kristensen, T. R., Bech-Hansen, D., & Elklit, A. (2019). The psychological subtype of intimate partner violence and its effect on mental health: protocol for a systematic review and meta-analysis. *Systematic reviews, 8*(1), 1-10.

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics, 125*, 37-46.

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*: Cambridge University Press.

Duluth Domestic Abuse Intervention Program. (1984). Power and Control Wheel. Retrieved from https://www.theduluthmodel.org/wp-content/uploads/2017/03/PowerandControl.pdf. Accessed 13th July 2022.

Dutton, M. A., Goodman, L. A., & Schmidt, R. J. (2005). *Development and validation of a coercive control measure for intimate partner violence: Final technical report*. Retrieved from https://www.ojp.gov/library/publications/development-and-validation-coercive-control-measure-intimate-partner-violence

Elzinga, P., Poelmans, J., Viaene, S., & Dedene, G. (2009). Detecting Domestic Violence: Showcasing a Knowledge Browser based on Formal Concept Analysis and Emergent Self Organizing Maps. *Semantic Scholar (Preprint)*, 11-18. doi:https://pdfs.semanticscholar.org/5d95/f126c7135a4fd8002a87701777df86a28dc6.pdf

European Union Agency for Fundamental Rights. (2014). *Violence against women: an EU-wide survey*. Retrieved from https://fra.europa.eu/sites/default/files/fra_uploads/fra-2014-vaw-survey-main-results-apr14_en.pdf

Evans, J. A., & Aceves, P. (2016). Machine translation: mining text for social theory. *Annual Review of Sociology, 42*, 21-50.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin, 76*(5), 378.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Rossi, F. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689-707.

Follingstad, D. R. (2007). Rethinking current approaches to psychological abuse: Conceptual and methodological issues. *Aggression and Violent Behavior, 12*(4), 439-458.

Follingstad, D. R. (2009). The impact of psychological aggression on women's mental health and behavior: The status of the field. *Trauma, Violence, & Abuse, 10*(3), 271-289.

Follingstad, D. R. (2011). A measure of severe psychological abuse normed on a nationally representative sample of adults. *JOURNAL OF INTERPERSONAL VIOLENCE, 26*(6), 1194-1214.

Follingstad, D. R., Coyne, S., & Gambone, L. (2005). A representative measure of psychological aggression and its severity. *Violence and Victims, 20*(1), 25-38.

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv., 51*(4), Article 85. doi:10.1145/3232676

Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., & Dell, N. (2018). *“A Stalker's Paradise” How Intimate Partner Abusers Exploit Technology.* Paper presented at the Proceedings of the 2018 CHI conference on human factors in computing systems.

Garrett, A., & Hassan, N. (2019). Understanding the silence of sexual harassment victims through the #WhyIDidntReport movement. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 649–652.

Gauthier, R. P., & Wallace, J. R. (2022). The Computational Thematic Analysis Toolkit. *Proceedings of the ACM on Human-Computer Interaction, 6*, 1-15. doi:https://doi.org/10.1145/3492844

Ge, Y., Guo, Y., Das, S., Al-Garadi, M. A., & Sarker, A. (2023). Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics, 144*, 104458. doi:https://doi.org/10.1016/j.jbi.2023.104458

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy, 9*(3), 330-338.

Graham-Kevan, N., & Archer, J. (2003). Physical aggression and control in heterosexual relationships: The effect of sampling. *Violence and Victims, 18*(2), 181.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology, 8*(1), 23.

Hartung, M., Klinger, R., Schmidtke, F., & Vogel, L. (2017). *Identifying right-wing extremism in German Twitter profiles: A classification approach.* Paper presented at the International conference on applications of natural language to information systems.

Heaton, J. (2016). *An empirical analysis of feature engineering for predictive modeling.* Paper presented at the SoutheastCon 2016.

Hinds, J., Parkhouse, T., & Hotchin, V. (2021). Assessing the quality of studies using machine learning for personality assessment: A systematic review. *PsyArxiv*. Retrieved from https://psyarxiv.com/4g8ec/download?format=pdf

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.

Hoffman, P. (1984). Psychological abuse of women by spouses and live-in lovers. *Women & Therapy, 3*(1), 37-49.

Homan, C., Schrading, J., Ptucha, R., Cerulli, C., & Alm, C. (2020). Quantitative Methods for Analyzing Intimate Partner Violence in Microblogs: Observational Study. *J Med Internet Res, 22*(11).

Home Office. (2021). Review of the controlling or coercive behaviour offence. Retrieved from https://www.gov.uk/government/publications/review-of-the-controlling-or-coercive-behaviour-offence/review-of-the-controlling-or-coercive-behaviour-offence#key-findings-and-research-recommendations. Accessed 25th May 2022.

Home Office. (2022). Draft controlling or coercive behaviour statutory guidance. Retrieved from https://www.gov.uk/government/consultations/controlling-or-coercive-behaviour-statutory-guidance/draft-controlling-or-coercive-behaviour-statutory-guidance-accessible. Accessed 25th May 2022.

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., . . . O'Cathain, A. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for information, 34*(4), 285-291.

Houston-Kolnik, J. D., & Vasquez, A. L. (2022). Cognitive Interviewing: Lessons Learned and Recommendations for Structured Interviews with Survivors of Crime. *Journal of family violence, 37*(2), 325-335.

Hwang, Y. I., Zheng, L., Karystianis, G., Gibbs, V., Sharp, K., & Butler, T. (2020). Domestic violence events involving autism: a text mining study of police records in New South Wales, 2005-2016. *Research in Autism Spectrum Disorders, 78*.

Johnson, M. P., & Leone, J. M. (2005). The differential effects of intimate terrorism and situational couple violence: Findings from the National Violence Against Women Survey. *Journal of family issues, 26*(3), 322-349.

Kapania, S., Taylor, A. S., & Wang, D. (2023). *A hunt for the Snark: Annotator Diversity in Data Practices.* Paper presented at the Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.

Karystianis, G., Adily, A., Schofield, P. W., Greenberg, D., Jorm, L., Nenadic, G., & Butler, T. (2019). Automated Analysis of Domestic Violence Police Reports to Explore Abuse Types and Victim Injuries: Text Mining Study. *J Med Internet Res, 21*(3), e13067. doi:10.2196/13067

Karystianis, G., Adily, A., Schofield, P. W., Wand, H., Lukmanjaya, W., Buchan, I., . . . Butler, T. (2022). Surveillance of Domestic Violence Using Text Mining Outputs From Australian Police Records. *Frontiers in Psychiatry, 12*.

Karystianis, G., Cabral, R. C., Han, S. C., Poon, J., & Butler, T. (2021). Utilizing Text Mining, Data Linkage and Deep Learning in Police and Health Records to Predict Future Offenses in Family and Domestic Violence. *Front Digit Health, 3*, 602683. doi:10.3389/fdgth.2021.602683

Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., & De Choudhury, M. (2021). A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1-34.

Kulesza, T., Amershi, S., Caruana, R., Fisher, D., & Charles, D. (2014). *Structured labeling for facilitating concept evolution in machine learning.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Lagdon, S., Armour, C., & Stringer, M. (2014). Adult experience of mental health outcomes as a result of intimate partner violence victimisation: a systematic review. *European journal of psychotraumatology, 5*(1), 24794.

Lagdon, S., Jordan, J.-A., Devine, P., Tully, M. A., Armour, C., & Shannon, C. (2022). Public Understanding of Coercive Control in Northern Ireland. *Journal of family violence, 38*(1), 39-50.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Lawrence, E., Yoon, J., Langer, A., & Ro, E. (2009). Is psychological aggression as detrimental as physical aggression? The independent effects of psychological aggression on depression and anxiety symptoms. *Violence and Victims, 24*(1), 20-35.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Gutmann, M. (2009). Computational social science. *Science, 323*(5915), 721-723.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240.

Li, C., Sheng, Y., Ge, J., & Luo, B. (2019). Apply event extraction techniques to the judicial field. *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 492–497. Retrieved from https://doi.org/10.1145/3341162.3345608

Liu, P., Li, W., & Zou, L. (2019). *NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers.* Paper presented at the Proceedings of the 13th international workshop on semantic evaluation.

Liu, Y., Li, Q., Liu, X., Zhang, Q., & Si, L. (2019). Sexual Harassment Story Classification and Key Information Identification. *Proceedings of the 28th ACM international conference on information and knowledge management*, 2385–2388.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory, 28*(2), 129-137.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Mahoney, J. C., Farrell, D. M., & Murphy, C. M. (2022). Prevalence and Predictors of Cyber Psychological Abuse among Adults. *Journal of family violence, 37*(1), 151-163.

Marshall, L. L. (1996). Psychological abuse of women: Six distinct clusters. *Journal of family violence, 11*(4), 379-409.

Marshall, L. L. (1999). Effects of men's subtle and overt psychological abuse on low-income women. *Violence and Victims, 14*(1), 69-88.

McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A, 379*(2207), 20200363. doi:https://doi.org/10.1098/rsta.2020.0363

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

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Mishra, P., Yannakoudakis, H., & Shutova, E. (2019). Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews, 4*(1), 1-9.

Møller, A. G., Dalsgaard, J. A., Pera, A., & Aiello, L. M. (2023). Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

More, K., & Francis, F. (2021). Analyzing the Impact of Domestic Violence on Social Media using Natural Language Processing. *Proceedings of the 2021 IEEE Pune Section International Conference (PuneCon)*, 1-5. doi:10.1109/PuneCon52575.2021.9686490

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). A BERT-based transfer learning approach for hate speech detection in online social media. *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications*, 928-940.

Muller, M., Wolf, C. T., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., . . . Duesterwald, E. (2021). *Designing ground truth and the social life of labels.* Paper presented at the Proceedings of the 2021 CHI conference on human factors in computing systems.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science, 43*(1), 88-102.

Norwood, A., & Murphy, C. (2012). What forms of abuse correlate with PTSD symptoms in partners of men being treated for intimate partner violence? *Psychological Trauma: Theory, Research, Practice, and Policy, 4*(6), 596.

Office for National Statistics. (2017). Crime Survey for England and Wales. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/domesticabusefindingsfromthecrimesurveyforenglandandwales/yearendingmarch2017. Accessed 9th December 2021.

Office for National Statistics. (2018a). *Domestic abuse: findings from the Crime Survey for England and Wales: year ending March 2018*. Retrieved from https://backup.ons.gov.uk/wp-content/uploads/sites/3/2018/11/Domestic-abuse-findings-from-the-Crime-Survey-for-England-and-Wales-year-ending-March-2018.pdf

Office for National Statistics. (2018b). Homicide in England and Wales: year ending March 2018. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/homicideinenglandandwales/yearendingmarch2018#how-are-victims-and-suspects-related. Accessed July 2022.

Oldrup, H., Andersen, S., Kjær, S., Nielsen, N., & von Rosen, C. (2018). Psykiske, fysiske og sociale konsekvenser af psykisk vold i parforhold – kortlægning af forskning. Lev Uden Vold. København. Retrieved from https://levudenvold.dk/wp-content/uploads/2018/08/rapport-konsekvenser-af-psykisk-vold-web.pdf.

ONS. (2020). Domestic abuse victim characteristics, England and Wales: year ending March 2020. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/domesticabusevictimcharacteristicsenglandandwales/yearendingmarch2020#sex. Accessed 8th Febuary 2022.

OpenAI. (2023). GPT-4 Technical Report. *ArXiv, abs/2303.08774*.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews, 5*(1), 1-10.

Øverlien, C., Hellevik, P. M., & Korkmaz, S. (2020). Young women's experiences of intimate partner violence–narratives of control, terror, and resistance. *Journal of family violence, 35*(8), 803-814.

Pandora Project. (2023). Cycle of Abuse. Retrieved from https://www.pandoraproject.org.uk/cycle-of-abuse/. Accessed 31st July 2023.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates.* Retrieved from www.LIWC.net.

Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation.* Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Pichaiah, S. (2023). Data Explainability: The Counterpart to Model Explainability. Retrieved from https://www.dataversity.net/data-explainability-the-counterpart-to-model-explainability/. Accessed 27th July 2023.

Pico-Alfonso, M. A., Garcia-Linares, M. I., Celda-Navarro, N., Blasco-Ros, C., Echeburúa, E., & Martinez, M. (2006). The impact of physical, psychological, and sexual intimate male partner violence on women's mental health: depressive symptoms, posttraumatic stress disorder, state anxiety, and suicide. *Journal of women's health, 15*(5), 599-611.

Poelmans, J., Elzinga, P., & Dedene, G. (2013). Retrieval of criminal trajectories with an FCA-based approach. *Proceedings of the FCAIR 2013 Formal Concept Analysis meets Information Retrieval workshop, co-located with the 35th European Conference on Information Retrieval (ECIR 2013), 977*, 83-94. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84922531785&partnerID=40&md5=be4d63208c0bcde7f74dfa931b6cdebf

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2008). An exploration into the power of formal concept analysis for domestic violence analysis. *Advances in Data Mining. Medical Applications, E-Commerce, Marketing, and Theoretical Aspects: 8th Industrial Conference, ICDM 2008 Leipzig, Germany, July 16-18, 2008, 8*, 404-416. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-48949103892&doi=10.1007%2f978-3-540-70720-2_31&partnerID=40&md5=36b48570ae29383dc7b36c7bf1ea7429

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2009). A case of using formal concept analysis in combination with emergent self organizing maps for detecting domestic violence. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5633*, 247-260. doi:10.1007/978-3-642-03067-3_20

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2010). Curbing domestic violence: instantiating C-K theory with formal concept analysis and emergent self-organizing maps. *Intelligent Systems in Accounting, Finance and Management, 17*(3), 167.

Poelmans, J., Elzinga, P., Viaene, S., & Dedene, G. (2011). Formally analysing the concepts of domestic violence. *Expert systems with applications, 38*(4), 3116-3130.

Poelmans, J., Elzinga, P., Viaene, S., Dedene, G., & Van Hulle, M. M. (2009). Analyzing domestic violence with topographic maps: A comparative study. *Advances in Self-Organizing Maps: 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 7*, 246-254.

Poelmans, J., Elzinga, P., Viaene, S., Hulle, M. M. V., & Dedene, G. (2009). How Emergent Self Organizing Maps Can Help Counter Domestic Violence. *2009 WRI World Congress on*

Computer Science and Information Engineering, 4, 126-136. doi:10.1109/CSIE.2009.299

Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M. M., & Dedene, G. (2009). Gaining insight in domestic violence with Emergent Self Organizing Maps. *Expert systems with applications, 36*(9), 11864-11874.

Poelmans, J., Van Hulle, M., Viaene, S., Elzinga, P., & Dedene, G. (2011). Text mining with emergent self organizing maps and multi-dimensional scaling: A comparative study on domestic violence. *Applied Soft Computing Journal, 11*(4), 3870-3876.

Porrúa García, C., Rodríguez Carballeira, Á., Escartín Solanelles, J., Gómez Benito, J., Almendros Rodríguez, C., & Martín Peña, J. (2016). Development and validation of the scale of psychological abuse in intimate partner violence (EAPA-P). *Psicothema*. Retrieved from http://hdl.handle.net/11162/118377

Prabakaran, S., Waylan, M., & Penfold, C. (2017). An Introduction to Machine Learning. Retrieved from https://bioinformatics-training.github.io/intro-machine-learning-2017/. Accessed 9th December 2021.

Prabhakaran, V., Davani, A. M., & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society, 7*(2). doi:10.1177/20563051211019004

Ramos, J. (2003). *Using tf-idf to determine word relevance in document queries.* Paper presented at the Proceedings of the first instructional conference on machine learning.

Reddit.com. Retrieved from Reddit.com.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). *" Why should i trust you?" Explaining the predictions of any classifier.* Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Rodriguez, M. Y., & Storer, H. (2020). A computational social science perspective on qualitative data exploration: Using topic models for the descriptive analysis of social media data. *Journal of Technology in Human Services, 38*(1), 54-86.

Rogers, M., Rumley, T., & Lovatt, G. (2019). The change up project: Using social norming theory with young people to address domestic abuse and promote healthy relationships. *Journal of family violence, 34*(6), 507-519.

Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018, 8-13 July 2018). *A "Deeper" Look at Detecting Cyberbullying in Social Networks.* Paper presented at the 2018 International Joint Conference on Neural Networks (IJCNN).

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., . . . Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior, 93*, 333-345.

Röttger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2021). Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. *arXiv preprint arXiv:2112.07475*.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys, 16*, 1-85.

Sackett, L. A., & Saunders, D. G. (1999). The impact of different forms of psychological abuse on battered women. *Violence and Victims, 14*(1), 105-117.

SafeLives. (2019). Psychological Violence. Retrieved from https://www.safelivesresearch.org.uk/Comms/Psychological%20Violence%20-%20Full%20Report.pdf. Accessed 27th July 2023.

Salawu, S., He, Y., & Lumsden, J. (2020). Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing, 11*(1), 3-24. doi:10.1109/TAFFC.2017.2761757

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Sanchez-Moya, A. (2017). Corpus-driven insights into the discourse of women survivors of Intimate Partner Violence. *QUADERNS DE FILOLOGIA-ESTUDIS LINGUISTICS, 22*, 215-243.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schmidt, A., & Wiegand, M. (2017). *A survey on hate speech detection using natural language processing.* Paper presented at the Proceedings of the fifth international workshop on natural language processing for social media.

Schrading, N., Alm, C. O., Ptucha, R., & Homan, C. M. (2015a). An analysis of domestic abuse discourse on reddit. *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2577-2583.

Schrading, N., Alm, C. O., Ptucha, R., & Homan, C. M. (2015b). WhyIStayed, #WhyILeft: Microblogging to make sense of domestic abuse. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1281-1286.

Shahi, N., Shahi, A. K., Phillips, R., Shirek, G., Lindberg, D. M., & Moulton, S. L. (2021). Using deep learning and natural language processing models to detect child physical abuse. *Journal of Pediatric Surgery, 56*(12), 2326-2332.

Siebert, J., Joeckel, L., Heidrich, J., Nakamichi, K., Ohashi, K., Namba, I., . . . Aoyama, M. (2020). *Towards guidelines for assessing qualities of machine learning systems.* Paper presented at the International Conference on the Quality of Information and Communications Technology.

SpaCy. (2023). SpaCy. Retrieved from https://spacy.io. Accessed 1st February 2024.

Stark, E. (2009). *Coercive control: The entrapment of women in personal life*: Oxford University Press.

Stark, E. (2013). Coercive control. *Violence against women: Current theory and practice in domestic abuse, sexual violence and exploitation*, 17-33.

Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., & Shakeel, H. (2019). Deep Learning for Multi-Class Identification From Domestic Violence Online Posts. *IEEE Access, 7*, 46210-46224.

Subramani, S., Vu, H. Q., & Wang, H. (2017). Intent Classification Using Feature Sets for Domestic Violence Discourse on Social Media. *2017 4th Asia-Pacific World Congress on Computer Science and Engineering*, 129-136.

Subramani, S., Wang, H., Islam, M., Ulhaq, A., & O'Connor, M. (2018). Child Abuse and Domestic Abuse: Content and Feature Analysis from Social Media Disclosures.

Databases Theory and Applications: 29th Australasian Database Conference, ADC 2018, Gold Coast, QLD, Australia, May 24-27, 39, 174-185.

Subramani, S., Wang, H., Vu, H. Q., & Li, G. (2018). Domestic Violence Crisis Identification From Facebook Posts Based on Deep Learning. *IEEE Access, 6*, 54075-54085.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to fine-tune bert for text classification?* Paper presented at the Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18.

Taft, C. T., Murphy, C. M., King, L. A., Dedeyn, J. M., & Musser, P. H. (2005). Posttraumatic stress disorder symptomatology among partners of men in treatment for relationship abuse. *Journal of abnormal psychology, 114*(2), 259.

Tanczer, L. M., López-Neira, I., & Parkin, S. (2021). 'I feel like we're really behind the game': perspectives of the United Kingdom's intimate partner violence support sector on the rise of technology-facilitated abuse. *Journal of gender-based violence, 5*(3), 431-450.

Thompson, M. P., Basile, K. C., Hertz, M. F., & Sitterle, D. (2006). *Measuring intimate partner violence victimization and perpetration; a compendium of assessment tools*: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Division of Violence Prevention.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Tolman, R. M. (1989). The development of a measure of psychological maltreatment of women by their male partners. *Violence and Victims, 4*(3), 159-177.

Torregrosa, J., Bello-Orgaz, G., Martinez-Camara, E., Del Ser, J., & Camacho, D. (2021). A survey on extremism analysis using natural language processing. *arXiv preprint arXiv:2104.04069*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Vatnar, S. K. B., & Bjørkly, S. (2008). An interactional perspective of intimate partner violence: An in-depth semi-structured interview of a representative sample of help-seeking women. *Journal of family violence, 23*(4), 265-279.

Victor, B., Perron, B., Sokol, R., Fedina, L., & Ryan, J. (2021). Automated Identification of Domestic Violence in Written Child Welfare Records: Leveraging Text Mining and Machine Learning to Enhance Social Work Research and Evaluation. *JOURNAL OF THE SOCIETY FOR SOCIAL WORK AND RESEARCH, 12*(4), 631-655.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur), 53*(3), 1-34.

Waqas, A., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLoS One, 14*(9), e0222194. doi:10.1371/journal.pone.0222194

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal, 32*(4), 791-813.

Wilson, M., Spike, E., Karystianis, G., & Butler, T. (2021). Nonfatal Strangulation During Domestic Violence Events in New South Wales: Prevalence and Characteristics Using

Text Mining Study of Police Narratives. *VIOLENCE AGAINST WOMEN, 28*(10), 2259-2285. doi:10.1177/10778012211025993

Withall, A., Karystianis, G., Duncan, D., Hwang, Y. I., Kidane, A. H., & Butler, T. (2022). Domestic Violence in Residential Care Facilities in New South Wales, Australia: A Text Mining Study. *Gerontologist, 62*(2), 223-231. doi:10.1093/geront/gnab068

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Funtowicz, M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1301.3781*.

Women's Aid. (2021a). How Common Is Domestic Abuse? Retrieved from https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/how-common-is-domestic-abuse/. Accessed 8th Febuary 2022.

Women's Aid. (2021b). What Is Domestic Abuse? Retrieved from https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/. Accessed 9th December 2021.

Women's Aid. (2023). Myths About Domestic Abuse. Retrieved from https://www.womensaid.org.uk/information-support/what-is-domestic-abuse/myths/. Accessed 27th July 2023.

Wood, L., Backes, B., Baumler, E., & McGiffert, M. (2021). Examining the impact of duration, connection, and dosage of domestic violence services on survivor well-being. *Journal of family violence, 37*, 221-233.

World Health Organisation. (2021). Violence Against Women: Key Facts. Retrieved from https://www.who.int/news-room/fact-sheets/detail/violence-against-women. Accessed 9th December 2021.

World Health Organisation. (2023). Violence Info: Intimate Partner Violence. Retrieved from https://apps.who.int/violence-info/intimate-partner-violence/. Accessed 30th July 2023.

Xie, N., Ras, G., van Gerven, M., & Doran, D. (2020). Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*.

Xu, H., Zeng, J., Tai, Z., & Hao, H. (2022). Public Attention and Sentiment toward Intimate Partner Violence Based on Weibo in China: A Text Mining Approach. *Healthcare (Basel), 10*(2). doi:10.3390/healthcare10020198

Xue, J., Chen, J., Chen, C., Hu, R., & Zhu, T. (2020). The Hidden Pandemic of Family Violence During COVID-19: Unsupervised Learning of Tweets. *J Med Internet Res, 22*(11).

Xue, J., Chen, J., & Gelles, R. (2019). Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter. *VIOLENCE AND GENDER, 6*(2), 105-114.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). *Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy.* Paper presented at the 2020 Conference on fairness, accountability, and transparency.

Zaman, A., Kautz, H., Silenzio, V., Hoque, M. E., Nichols-Hadeed, C., & Cerulli, C. (2021). Discovering intimate partner violence from web search history. *Smart Health, 19*. Retrieved from https://doi.org/10.1016/j.smhl.2020.100161

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111-151.

Zhang, Z. (2018). *Improved adam optimizer for deep neural networks.* Paper presented at the 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS).

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). *A heuristic approach to determine an appropriate number of topics in topic modeling*. Paper presented at the BMC bioinformatics.