Extracting Knowledge from Complex Unstructured Corpora: Text Classification and a Case Study on the Safeguarding Domain

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Aleksandra I. Edwards

January 2022

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Dedicated to Elizabeth, Tom, Jeni, and Penka
Their love and support will always be cherished.

Abstract

The advances in internet, data collection and sharing technologies have lead to an increase in the amount of unstructured information in the form of news, articles, and social media. Additionally, many specialised domains such as the medical, law, and social science-related domains use unstructured documents as a main platform for collecting, storing and sharing domain-specific knowledge. However, the manual processing of these documents is a resource-consuming and error-prone process. This is especially apparent when the volume of the documents that need annotating constantly increases over time. Therefore, automated information extraction techniques have been widely used to efficiently analyse text and discover patterns. Specifically, text classification methods have become valuable for specialised domains for organising content, such as patient notes, and help fast topic-based retrieval of information. However, many specialised domains suffer from lack of data and class imbalance problems because documents are hard to obtain. In addition, the manual annotation needs to be performed by experts which can be costly. This makes the application of supervised classification approaches a challenging task.

In this thesis, we research methods for improving the performance of text classifiers for specialised domains with limited amounts of data and highly domain-specific terminology where the annotation of documents is performed by domain experts. First, we study the applicability of traditional feature enhancement approaches using publicly available resources for improving classifiers performance for specialised domains. Then, we conduct extensive research into suitability of existing classification algorithms

and the importance of both domain and task specific data for few-shot classification which helps identify classification strategies applicable to small datasets. This gives the basis for the development of a methodology for improving a classifier's performance for few-shot settings using text generation-based data augmentation techniques. Specifically, we aim to improve quality of generated data by using strategies for selecting class representative samples from the original dataset used to produce additional training instances. We perform extensive analysis, considering multiple strategies, datasets, and few-shot text classification settings.

Our study uses a corpus of safeguarding reports as an exemplary case study of a specialised domain with a small volume of data. The safeguarding reports contain valuable information about learning experiences and reflections on tackling serious crimes involving children and vulnerable adults. They carry great potential to improve multi-agency work and help develop better crime prevention strategies. However, the lack of centralised access and the constant growth of the collection, make the manual analysis of the reports unfeasible. Therefore, we collaborated with the Crime and Security Research Institute (CSRI) at Cardiff University for the creation of a Wales Safeguarding Repository (WSR) for providing a centralised access to the safeguarding reports and means for automatic information extraction. The aim of the repository is to facilitate efficient searchability of the collection and thus help free up resources and assist practitioners from health and social care agencies in making faster and more accurate decisions. In particular, we apply methods identified in the thesis, in order to support automated annotation of the documents using a thematic framework, created by subject-matter experts. Our close work with domain experts throughout the thesis allowed incorporating experts‘ knowledge into classification and augmentation techniques which proved beneficial for the improvement of automated supervised methods for specialised domains.

Acknowledgements

I would like to give thanks my supervisors Alun Preece and Hélène De Ribaupierre for their support, encouragement and the tremendous amount of help throughout these years. Further, I would like to give additional thanks to Jose Camacho-Collados for his valuable input over the last two years of my PhD for his guidance that helped better shape this thesis. I am incredibly thankful and grateful for knowing these three people, especially Alun Preece whom guided this thesis to completion.

My gratitude also goes to the team at the Crime and Security Research Institute and especially the people part of the WSR project for their valuable input on the various stages of my PhD. I'm forever thankful for their warmth, friendliness and readiness to help. I'd also like to thank my friends in COMSC and non-academic friends for the nice memories we've created throughout these years and for their support and advice in difficult moments.

I would also like to give a special thank you to Tom, Mo and Luke for their help in the final hours of this thesis. The time that they spent helping me will always be cherished.

Finally, I am deeply grateful to my family for their love, belief in me, encouragement and sacrifices they have made for me throughout my life. Without their constant support, I would not have been able to work on the PhD and finalise it. They deserve my utmost thanks.

Funding Acknowledgements

This PhD was funded by the School of Computer Science and Informatics at Cardiff University. The PhD was conducted in collaboration with the Wales Safeguarding Repository (WSR) project, funded by the National Independent Safeguarding Board (NISB), the Crime and Security Research Institute at Cardiff University (CSRI), and the School of Social Sciences at Cardiff University (SOCSI). The first stage of development of WSR took place between March 2018 — May 2019 and a second stage of development commenced in May 2021.

Contents

List of Publications

The work introduced in this thesis is based on the following publications.

- Edwards et al. [2019] — Aleksandra Edwards, Alun Preece, and Hélène De Ribaupierre. *Knowledge extraction from a small corpus of unstructured safeguarding reports.* In *European Semantic Web Conference*, pages 38-42, Portoroź, Slovenia, 2019. *Springer*. — This work is the basis for Section 3.2 from Chapter 3. Contributions include: methodology, analysis, paper draft.

- Edwards et al. [2020] — Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. *Go simple and pre-train on domain-specific corpora: On the role of training data for text classification*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522-5529, 2020. — This work is the basis for Chapter 5, except analysis performed for the Safeguarding domain. Contributions include: methodology, analysis, paper draft.

- Edwards et al. [2021a] — Aleksandra Edwards, David Rogers, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. *Predicting themes within complex unstructured texts:A case study on safeguarding report*. In *2nd International Workshop Deep Learning meets Ontologies and Natural Language Processing, European Semantic Web Conference (ESWC 2021)*, 2021. — This work is the basis for Sections 4.2, 4.3, 4.4, 4.5 from Chapter 4. Contributions include: methodology, analysis, paper draft.

- Edwards et al. [2021b] — Aleksandra Edwards, Asahi Ushio, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. *Guiding generative pre-trained language models fordata augmentation in few-shot text classification*. Submitted to *the 2021 Conference on Emirical Methods in Natural Language Processing (EMNLP 2021)*, 2021. — This work is the basis for Chapter 6. Contributions include: methodology, analysis and development on text generation-based data augmentation without development of baseline approaches, paper draft.

List of Figures

List of Tables

List of Acronyms

APRs Adult Practice Reviews

BERT Bidirectional Encoder Representations from Transformers

BOW Bag-of-Words

CBOW Continuous Bag of Words Model

CNN Convolutional Neural Network

CPRs Child Practice Reviews

CSRI Crime and Security Research Institute

DA Data Augmentation

DAPT Domain-adaptive pre-training

DHRs Domestic Homicide Reports

GANs Generative Adversarial Networks

GNB Gaussian Naive Bayes

HCI Human Computer Interaction

IE Information Extraction

LAMBADA Language Model-Based Data Augmentation

LDA Latent Dirichlet Allocation

LG Logistic Regression

LSTM Long-Short Term Memory Neural Network

MHHRs Mental Health Homicide Reviews

ML Machine Learning

MLM Masked Language Model

NER Named Entity Extraction

NLP Natural Language Processing

OOV out-of-vocabulary

RNN Recursive Neural Network

RF Random Forest

SR Sentence replacement-based

SVM Support Vector Machines

TA Text Augmentation

TAPT Task-adaptive pre-training

TG Text Generation-based

uSIF unsupervised Smooth Inverse Frequency

FSL Few-Shot Learning

WR Word replacement-based

WSR Welsh Safeguarding Repository

ZSL Zero-Shot Learning

Chapter 1

Introduction

A large percentage of corporate information exists in textual and unstructured format [Aggarwal and Zhai, 2012]. Further, the advances in internet technologies lead to significant increase in the amount of unstructured information in the form of news, articles, and social media [Singh, 2018]. The unstructured texts contain valuable knowledge, however, manually processing these large volumes of data to identify useful patterns is a time-consuming, resource-expensive and error-prone process [Singh, 2018]. This leads to the need for automated text analysis methods to support knowledge extraction [Hu et al., 2019, Ali, 2019, Türker et al., 2019, Metzler et al., 2016 (accessed February 3, 2014, Bernard and Bernard, 2013, Sinoara et al., 2019, Singh, 2018]. Information Extraction (IE) and text classification techniques are widely used to efficiently analyse free text and to discover valuable patterns within unstructured corpora [Singh, 2018]. IE and classification methods are successfully used for many applications such as improving customer services where companies use automated tools to retrieve, structure, and classify relevant customer information. In this thesis, we specifically focus on the task of using supervised approaches for classifying information. Text classification is a widely researched problem as it has a wide range of applicability in many domains and tasks. It can be defined as a process using supervised machine learning techniques in order to assign one or more class labels or categories from a predefined set of labels or categories to a given text, according to its content [Deng et al., 2019, Kong et al., 2019, Zhong and Enke, 2019]. Text classification techniques are extensively used in web-based information retrieval systems for classifying web pages and news, recommend-

ation systems for suggesting items to users based on the user's interests [Aggarwal, 2016], and for information filtering such as spam email filtering [Deng et al., 2019, Aggarwal and Zhai, 2012].

Text classification has also become increasingly valuable for more specialised domains such as medicine, social sciences, healthcare, psychology, and law [Kowsari et al., 2019]. For instance, most textual information in the medical domain is presented in an unstructured or narrative form with ambiguous terms and typographical errors. Such information needs to be available instantly throughout the patient-physician encounters in different stages of diagnosis and treatment [Lauría and March, 2011]. Therefore, many automated text classification approaches are widely applied for organising electronic health record (EHR) data [Zhang et al., 2018]. Further, in the social science domain, text classification has increasingly been applied to understanding human behavior [Nobles et al., 2018, Ofoghi and Verspoor, 2017]. However, many specialised domains suffer from data scarcity and class imbalance problems [Türker et al., 2019, Zhang and Wu, 2015, Shams, 2014, Kumar et al., 2020] because documents are hard to obtain and costly to annotate as experts are required.

State-of-the-art approaches in text classification are based on using neural network (NN) models, and especially language models, pre-trained using publicly available text documents [Gururangan et al., 2020, Rogers et al., 2020]. However, NN models require large computational resources that are not always available and have important environmental implications [Strubell et al., 2019]. Further, to train a NN model usually requires a large volume of manually annotated data, which greatly limits the practicality and scalability of models [Lyu et al., 2020, Yang et al., 2020, Strubell et al., 2019, Sainz and Rigau, 2021, Christopher et al., 2008, Sebastiani, 2002, Lewis et al., 2004, Lyu et al., 2020, Türker, 2019, Li and Yang, 2018, Cawley and Talbot, 2010, Colace et al., 2014] especially for domains with scarce amount of data.

Few-shot text classification is a widely used approach for addressing the problem of data sparsity and scarcity [Gupta et al., 2020] where it refers to the process of learn-

ing classifiers given only a few labeled examples (usually less than 20 per label) of each class [Wang et al., 2020]. Therefore, it is especially valuable for situations where labelled instances are hard or impossible to acquire due to privacy or ethical issues, rare occurrences of events, and the need of expert annotators. This method uses prior knowledge to generalise to new tasks using only a handful of labelled instances [Bailey and Chopra, 2018]. Many few-shot learning approaches consist of adapting pre-trained models, usually transformer-based models, to the domains or task at hand in order to improve performance of models for the given purposes [Gururangan et al., 2020, Gupta et al., 2020, Vaswani et al., 2017]. However, the applicability of pre-trained language models to domains with specialised terminology, especially with limited amount of unlabelled data, is not extensively researched area. Further, studies are limited in datasets and models used for analysing the use of recent language models and generative models over few-shot classification tasks. There is also lack of comparison between more traditional but less data-consuming approaches and recent language models and generative models over few-shot classification tasks.

Another approach for improving classifiers performance is enriching feature vectors using publicly available lexical resources [Faruqui et al., 2015, Mrkšić et al., 2017, Choi et al., 2017, Min et al., 2017, Salguero et al., 2018, Gazzotti et al., 2019]. This approach is widely used, especially in the medical domain, due to its simplicity and less resource-consuming nature compared to NN-based methods. However, a main drawback of these approaches is the need for the lexical resource to fit the needs of the domain and the classification task. Otherwise, there is need to develop domain-specific knowledge graphs. However, the creation of domain-specific knowledge graphs can be a time- and cost- consuming process as it requires the supervision of domain experts. Further, the created knowledge graph is often only applicable to the domain it was created for, especially when the domain is specialised and contains a high number of polysemous words.

The significant need for establishing supervised text classification approaches which

can be used for specialised domains with limited data is the main motivation for this thesis. We specifically focus on performing extensive comparison between existing classification strategies for various domains with limited data and we also look at how recent generative neural models can be used for enhancing few-shot classification performance where we focus on scenarios with less than 20 labelled instances per label.

1.1 Motivation

Throughout this thesis, we focus on a scenario motivated by collaboration with the Crime and Security Research Institute (CSRI), part of Cardiff University. The institute provides interdisciplinary expertise in crime and security research. Its aim is to help reduce crime and increase security by identifying problems and providing practical, well-researched solutions, that directly inform policy and practice. The institute has been running for six years and holds PhD students and researchers from different knowledge areas, studying and devising methodologies that facilitate the work of police, health, and social care agencies and thus can help reduce and prevent crime.

There is a variety of research projects currently underway in the institute related to police science, security and defence applications of computer science, alcohol and violence-related harm reduction, and digital behavioural analytics. One project that has been running since 2018 in CSRI is the Wales Safeguarding Repository (WSR) project. It involves a collaborative work between social scientists and computer scientists and aims to build a document repository for housing safeguarding reviews and reports. ‘Safeguarding’ is a term used to denote ‘measures to protect the health, well-being and human rights of individuals, which allow people, especially children, young adults, and vulnerable adults to live free from abuse, harm and neglect’ [Quality Commission, 2014].

Each safeguarding report contains key information about learning experiences and reflections on tackling serious incidents. The purpose of a safeguarding report is to

identify and describe related events that precede a serious safeguarding incident, and to reflect on agencies roles and the application of current practices. The reports carry great potential to improve multi-agency work and help develop better safeguarding practices and strategies. However, each report is lengthy and complex. Further, the current collection of 27 reports is expected to grow significantly in the near future with the addition of 500 historical reports. This makes the manual extraction of information a time-consuming and potentially bias-prone process. Therefore, a computer science team has been working alongside the sociology researchers since 2018 conducting NLP-related research on the data and implementing search tools developed as a result of the work in the thesis.

The social science part of the project is responsible for the development of a robust coding framework for highlighting themes within the reports. The thematic framework was developed to help identify common problems and issues in multi-agency work across different reports and it resulted in collaborative work between multiple subject-matter experts following standard approaches of performing thematic analysis in social science domain. The initial framework was heavily influenced by the findings of a thematic review looking across several safeguarding report types, presented by Robinson et al. [2019]. In this context, a *theme* refers to a main topic of discussion related to safeguarding incidents, specifically relevant to domestic homicide and mental health homicide. The need for experts to be able to search through the growing collection of the reports in line with the themes forms the knowledge-driven basis for classification.

All these make the safeguarding collection a good representative of domains associated with a small number of documents that are rich in specialised terminology and require annotation to be performed by subject-matter experts. Therefore, we use the safeguarding reports as a case study for testing our hypothesis in text classification.

Additionally, providing tools which automatically annotate the new reports can facilitate efficient browsing of the collection and improve access by practitioners. Further,

the automatic identification of similar documents based on their relevance to themes can help the discovery of common trends across the reviews and enable faster and more accurate decision-making by practitioners from health and social care agencies.

1.2 Contributions

The key theme and motivation behind the work it involves is:

How can the performance of text classifiers be improved for specialised domains with small corpus?

The thesis addresses this question through identifying text classification approaches that fit the needs of specialised domains with small corpus, with a focus on how the performance of these approaches can be improved using data augmentation techniques based on text generation and seed selection strategies. Previous work on improving and adapting classification models for limited training data discussed in Sections 2.2, 2.3.6, 2.3.7 from Background chapter are limited in scale, lack evaluation between different classification approaches, and do not consider domains with limited both labelled and unlabelled dataset. Further, recent research on using text generation techniques as part of data augmentation approaches (discussed in Section 2.3.8 from Background chapter) have shown the potential of these methods for improving text classification. However, previous work does not address the problems of data quality associated with text generation approaches. The central point addressed by this research is that incorporating expert knowledge into guiding large pre-trained language models can be beneficial for performing classification for specialised domains. In this work, the following research questions help illustrate the steps towards realising this thesis:

- **RQ 1:** Can publicly available lexical resources be used to support supervised learning for specialised domains?

- **RQ 2:** Which classification approaches help preserve subject-matter expert knowledge for annotating specialised unstructured texts, compared to human annotators?
- **RQ 3:** What are the most efficient approaches for few-shot classification in general and for specialised domains?
- **RQ 4:** Can text classification performance be improved through the use of data augmentation techniques based on text generation and seed selection strategies in few-shot settings in general and for specialised domains?

The main contributions made in this research work are outlined below.

- **Contribution 1:** Analysis into the use of traditional IE tools and semantic enrichment methods based on lexical resources for classification allows us to identify the challenges in extracting patterns from specialised texts. It also helps form foundations for further work towards substantiating the thesis with the knowledge that specialised texts require more context-aware methods for supervised learning tasks. The work relevant to this contribution is presented in **Chapter 3**.
- **Contribution 2:** A comparison between multiple classification approaches for extracting themes from the safeguarding reports allowed us to identify feature extraction, feature integration, and classification algorithms that are suitable for small specialised collections. Further, analysis comparing classifiers against expert annotators showed the potential of fine-tuned contextual models to preserve the knowledge of initial expert annotators but also introduced new questions regarding the suitability of state-of-the-art models for few-shot classification. The work relevant to this contribution is presented in **Chapter 4**.
- **Contribution 3:** Quantitative analysis covering four domains and six classification tasks into the role of training and unlabelled data for supervised text classification have been conducted. We performed analysis on both few-shot scenarios

with a balanced set and by randomly sampling different sized subsets from the original labelled datasets. The analysis revealed that in settings with small training data, regardless of the domain or task, a simple linear classifier coupled with domain-specific word embeddings appear to be more efficient than a more data-consuming language model, even when it is pre-trained on domain-relevant data. The work relevant to this contribution is presented in **Chapter 5**.

- **Contribution 4:** A data augmentation methodology using text generation techniques and seed selection strategies has been created for improving the quality of generated artificial sequences and subsequently classifier's performance in few-shot settings. Specifically, the seed selection strategies developed in the thesis have not been explored in previous research. We compare four seed selection strategies, including random selection, and two methods for fine-tuning text generation models for the classification task. Evaluation has been performed for three domains, four few-shot settings, and four baseline data augmentation methods. In general, the highest results were achieved when the text generation model is fine-tuned per label, even using only handful of instances, compared to the same model fine-tuned on the entire dataset. This shows the importance of label preservation techniques in the performance of text generation-based data augmentation methods. Additionally, seed selection strategies applied to domains closer to the datasets used for pre-training text generation models, led to classification improvements over random seed selection only when larger number of seed samples is selected. However, seed selection strategies for specialised texts, especially when incorporating expert knowledge, proved highly beneficial for few-shot text classification outperforming baselines and random seed selection methods. The work relevant to this contribution is presented in **Chapter 6**.

1.3 Thesis Structure

The chapters containing the remainder of this thesis are laid out as follows.

- **Chapter 2: Background and Research Domain** — This chapter introduces information extraction and text classification, and related work on classification approaches and data augmentation methods. Chapter 2 identifies gaps in literature related to classification approaches for scarce and specialised data and data augmentation methods for improving classification performance.

- **Chapter 3: Case Study and Exploratory Work: Traditional Information Extraction** — This chapter describes the case study of the WSR and safeguarding reports. It also presents initial work on the use of traditional methods for extracting information from texts with specialised terminology. The chapter shows that methods, based on standard IE tools, publicly available lexical resources, and statistical classifiers, are unsuitable for specialised domains such as the safeguarding and more contextually-aware approaches are needed. The work in this chapter relates to **Contribution 1**.

- **Chapter 4: Evaluation of State-of-the-art Classification Methods for the Safeguarding Domain** — This chapter analyses the suitability of state-of-the-art classification approaches for small collections of domain-specific texts. The chapter shows that state-of-the-art fine-tuned models do perform equally well to expert annotators for complex tasks when enough training data is provided. However, for limited amount of data, state-of-the-art approaches were outperformed by simpler linear classifier coupled with domain-adapted word embeddings. The work in this chapter relates to **Contribution 2**.

- **Chapter 5: Suitability of Text Classification Approaches for Few-shot Setting** — This chapter investigates the role of labelled and unlabelled data over the performance of supervised text classification tasks. The chapter shows that

using unlabelled domain-specific corpus, even if small, for training word embeddings or initialising language model improves performance significantly. Further, it shows that a simple linear classifier such as fastText coupled with domain-specific word embeddings is more suitable for limited training data then BERT, even when trained on domain-specific corpus. The work in this chapter relates to **Contribution 3**.

- **Chapter 6: Text Generation-based data Augmentation Techniques for Few-shot Text Classification** — This chapter describes approaches for improving text classification in a few-shot scenarios based on data augmentation using text generation techniques and seed selection strategies. The chapter shows that fine-tuning GPT-2 in a handful of label instances leads to consistent classification improvements and outperform competitive baselines. Further, it shows that guiding the generative process using domain expert seed selection can lead to further improvements. The work in this chapter relates to **Contribution 4**.

- **Chapter 7: Conclusions and Future Work** — This chapter concludes this thesis and summarises our contributions and findings. It also highlights work that could be undertaken to take this project further and covers current and future plans for developing on WSR.

Chapter 2

Background and Research Domain

As discussed in Chapter 1, text classification finds a high usage in many domains and applications. It is a widely researched area in NLP where a main focus has been the development of state-of-the-art neural models which has proven to give high results for many big data related tasks. However, current approaches are limited in researching the quality of developed techniques in domains with limited data and highly-domain specific terminology. The problem of applying supervised text classification in domains with limited data and specialised language is the basis of the research in this thesis. This chapter provides a review of main techniques and concepts in IE and text classification including a survey of some of the most relevant works of the area. Further, we point to gaps in the literature in relation to our problem focus set out in Chapter 1, i.e., unstructured specialised corpora with limited available data.

The increased information and documents load leads to the need for automated text analysis methods to support knowledge extraction [Hu et al., 2019, Ali, 2019, Türker et al., 2019, Metzler et al., 2016 (accessed February 3, 2014, Bernard and Bernard, 2013, Sinoara et al., 2019].Text classification applications, as well as text sources, are diverse. Examples of text classification applications are e-mail classification and spam filtering, news and scientific articles organization, financial forecasting, sentiment analysis, opinion mining, and topic labeling [Türker et al., 2019, Sinoara et al., 2019]. These applications can be represented as either sentence or text classification problems [Sinoara et al., 2019, Türker et al., 2019] where sentence classification refers to the process of assigning labels to single sentences of the corpus while text classifica-

tion refers to the process of assigning labels to sequences longer than a sentence (i.e., a section, paragraph, or an entire document).

2.1 Information Extraction

IE technologies help to efficiently and effectively analyse free text and to discover valuable and relevant knowledge from it in the form of structured information [Singh, 2018]. Hence, IE refers to the use of computational methods to identify relevant pieces of information in document generated for human use and convert this information into a representation suitable for computer based storage, processing, and retrieval [Singh, 2018]. IE sub-tasks include Named Entity Recognition (NER), co-reference resolution, named entity linking, relation extraction, knowledge base reasoning [Singh, 2018, Tang et al., 2008]. These tasks are part of many NLP applications such as machine translation, question answering, natural language understanding, and text summarisation [Singh, 2018, Tang et al., 2008]. IE methods are used for improving customer services where companies use tools to retrieve, structure, and classify relevant customer information. IE is also used in business analytics and business intelligence for acquiring market information and target advertising [Singh, 2018, Tang et al., 2008].

2.1.1 Traditional Information Extraction Approaches

We refer to traditional IE approaches to approaches which do not involve deep learning strategies. Traditional IE approaches are still widely used as they are well established in many fields and do not require large computational resources. These approaches can be divided into three groups [Singh, 2018, Tang et al., 2008, Chau et al., 2002] which are explained in the rest of this section.

Dictionary-based Approaches

Many traditional information extraction systems rely on the use of lexical resources such as dictionaries for extracting knowledge from unstructured text [Gentile et al., 2019, Tang et al., 2008]. A main challenge of using this approach is preparing complete and accurate gazetteers and keeping them up to date with the evolvement of the given domain [Kuriki et al., 2017].

Rule-based Approaches

Rule-based methods use human created heuristics to extract information from text. These rule based systems have been mostly used in information extraction for semi-structured web pages [Tang et al., 2008]. Very often rule-based approaches are used in combination with dictionaries. Similarly to the dictionary-based approaches, manually created rules might often need updates as data collections expand. Further, creating rules manually is a time- and resource-consuming process.

Statistical Machine Learning

These IE methods use supervised machine learning techniques for extracting knowledge. Machine Learning (ML) algorithms automatically learn the IE patterns by generalising from a given set of examples, rather than creating rules manually [Singh, 2018, Sugiyama, 2015]. Traditional ML algorithms such as Decision Trees [Quinlan, 1986], Naive Bayes classifier [McCallum and Nigam, 1998], Support Vector Machine (SVM) [Cortes and Vapnik, 1995], Conditional Random Fields (CRF) [Lafferty et al., 2001], Maximum Entropy (MaxEnt) [Jaynes, 1957] use features such as word frequencies for building vector representations for text used as training data.

Public Information Extraction Libraries

There are various publicly available IE libraries based on the traditional approaches, described above that have been widely used for extracting patterns of information for many domains and tasks. Examples include GATE [Cunningham et al., 2013], Stanford Core NLP [Manning et al., 2014], and UIMA [Kluegl et al., 2016]. We describe below libraries which we use in our initial analysis because they are built using different strategies and datasets which allow us to identify suitability of the different traditional IE methods for extracting knowledge from more specialised datasets such as the safeguarding reports.

Stanford Core NLP [Manning et al., 2014]: Stanford Core NLP is a set of human language technology tools and supports both NER and assigning sentiments to given text. It is based on a general implementation of (arbitrary order) linear chain CRF [Finkel et al., 2005] sequence models. Named entities are recognized using a combination of three CRF sequence taggers trained on Reuters newswire articles and emails containing seminar announcements at Carnegie Mellon University [Finkel et al., 2005]. Regarding sentiment analysis, Stanford Core NLP uses deep learning approach based on Recursive Neural Tensor Network. The neural model is trained on a corpus of movie reviews.

GATE [Cunningham et al., 2013]: GATE is a family of open source text analysis tools, which similarly to Stanford Core NLP supports NER tasks and sentiment analysis. It uses A Nearly-New IE system (ANNIE) for identifying named entities in text. In contrast to Stanford CoreNLP, it is using a rule-based approach rather than machine learning. It relies on a finite state algorithms, grammar rules (JAPE language) and gazetteers. The role of the gazetteer is to identify entity names in the text based on lists. The gazetteers have been created using news sources and articles. For our experiments, we use the default settings of ANNIE and we do not define JAPE rules.

Similarly, to the NER approach, GATE uses rules for performing sentiment analysis. It finds sentiment-containing words in linguistic relation with terms or entities. It consists of multiple dictionaries split into categories of negative, and positive emotions. Dictionaries give a starting score for sentiment words. It uses a number of linguistic sub-components to deal with issues such as negatives, adverbial modification, swear words, conditionals, sarcasm, etc. It consists of sentiment gazetteers, developed from sentiment words in WordNet. They have a starting 'strength' score which are modified by context words — adverbs, swear words, negatives, and so on.

Natural Language Toolkit (NLTK) [Bird et al., 2009]: NLTK is a python library for performing NLP analysis and it uses Maximum Entropy classifier for identifying named entities. It does not perform sentiment analysis.

SentiStrength [Thelwall et al., 2010]: SentiStrength estimates the strength of positive and negative sentiment in short informal texts, for informal language. It has human-level accuracy for short social web texts in English, except political texts. SentiStrength reports two sentiment strengths: -1 (not negative) to -5 (extremely negative) and 1 (not positive) to 5 (extremely positive). It is developed using an initial set of 2,600 MySpace classifications. The core of the algorithm consists of a lookup table of term sentiment strengths optimised by machine learning. The lookup table consists of 298 positive terms and 465 negative terms classified for either positive or negative sentiment strength. The emotion strength is specific to the contexts in which the words tend to be used in MySpace. The default manual word strengths are modified by a training algorithm to optimise the sentiment word strengths. This algorithm starts with the baseline human-allocated term strengths for the predefined list and then for each term assesses whether an increase or decrease of the strength by 1 would increase the accuracy of the classifications.

Google Cloud API `https://cloud.google.com/apis/`: Google Cloud API provides programming interfaces to Google Cloud Services such as NER and sentiment analysis. Google Sentiment Analysis inspects the given text and identifies the prevailing emotional opinion within the text. It determines a text sentiment as positive, negative, or neutral. Sentiment analysis attempts to determine the overall attitude (positive or negative) and is represented by numerical score and magnitude value. Despite Stanford Core NLP and Google Cloud API libraries using deep learning approaches, we describe them as part of traditional IE libraries as they are widely accepted for performing common IE tasks such as NER as well as sentiment analysis. The comparison between the aforementioned libraries presented here help analysis in Chapter 3.

2.1.2 Modern Information Extraction Approaches

2.1.3 Definition

Recently a shift occurred from more traditional rule-based IE and statistical-based ML approaches towards neural network-style ML approaches [Goldberg, 2016, Goodfellow et al., 2016]. Neural models can learn complex relationships which makes them a preferable method for many NLP tasks such as sentiment analysis or question answering, NER, and POS tagging [Sun et al., 2019]. Neural network-based approaches have shown great success in various applications such as object recognition [Krizhevsky et al., 2017] and speech recognition [Dahl et al., 2011]. Furthermore, many recent works showed that neural networks can be successfully used in a number of tasks in NLP. These include, but are not limited to, language modeling [Bengio et al., 2003], paraphrase detection [Socher et al., 2011] and word embedding extraction [Mikolov et al., 2013b].

Early Neural Models

Artificial neural networks are machine learning systems conceptually and structurally inspired by the brain neural system [Chen et al., 2019, Suk, 2017]. Earlier types of neural models are sequence-to-sequence models which transform a given sequence of elements, such as the sequence of words in a sentence, into another sequence. For instance, Recurrent Neural Network (RNN) are feed-forward NN, which process text in a sequential manner where sentences need to be processed word by word. They are designed for sequential data like text sentences, time-series, and other discrete sequences like biological sequences [Medsker and Jain, 2001, Suk, 2017, Aggarwal et al., 2018]. In essence, all these RNN types process sequential information by recurrence. Previous input is represented as the hidden state of the recurrent computation and each new input is processed and combined with the hidden state [Merkx and Frank, 2020].

RNNs have been firmly established as state-of-the-art approaches in sequence modeling and transduction problems such as language modeling and machine translation [Bahdanau et al., 2015, Cho et al., 2014, Sutskever et al., 2014, Mikolov et al., 2010]. The problem with RNN is that they process text from left-to-right or right-to-left and thus they remember dependencies only between contiguous words. Thus, they can learn only short-term dependencies. Long-short term memory neural models (LSTMs) are an extension to RNNs and they address this problem by introducing a feedback loop which helps learn long-term dependencies [Hochreiter and Schmidhuber, 1997] and thus help gain more contextual knowledge. Another widely used NN is Convolutional Neural Network (CNN) [Gehring et al., 2017] which was designed to work for processing images, however, it has proven to be suitable for textual data as well [Kalchbrenner et al., 2014]. CNNs allow processing in text to be done in parallel as each word on the input can be processed at the same time and does not necessarily depend on the previous words to be translated. However, CNN ignores dependencies between words in a sequence. In particular, LSTMs, sometimes in combinations with CNNs achieve good results for text classification [Xiao and Cho, 2016, Pilehvar et al.,

2017],and enable capturing long-range dependencies in a sequential manner. However, the sequential nature of both RNNs and LSTMs makes computation expensive [Merity et al., 2018, Yang et al., 2017, Vaswani et al., 2017], which limits the use of such models in practice [Wang et al., 2019c].

2.1.4 Transformer Models

In 2017, a new neural network architecture was introduced called Transformer [Vaswani et al., 2017]. The Transformer is fundamentally different from earlier neural models such as RNNs and LSTMs where the representation of each word is dependent on the representation of the previous word [Merkx and Frank, 2020]. Instead, the Transformer consists of self-attention layers Luong et al. [2015], allowing to 'attend' to parts of previous input directly and thus data can be processed in a non-sequential manner where each word representation is connected to the representation of every other word in the sequence, rather than processing words one-by-one [Merkx and Frank, 2020] (see Figure 2.1). This mechanism allows for much more parallelisation than RNNs and therefore reduced training times [Vaswani et al., 2017]. Additionally, the self-attention technique and the positional embeddings provide more information for the relationships between words and thus allow more contextual representation of text [Vaswani et al., 2017]. The reduced training time and the ability to capture long-range sequence features allow for Transformer models to reach a new state-of-the-art performance on several NLP tasks [Wolf et al., 2020, Devlin et al., 2019, Hayashi et al., 2019, Karita et al., 2019]. The Transformer has rapidly become the dominant architecture for natural language processing surpassing alternative neural models such as CNNs and RNNs in performance for tasks in both natural language understanding and natural language generation [Wolf et al., 2020]. The Transformer architecture is particularly suitable for pre-training on large text corpora [Wolf et al., 2020], leading to major gains in accuracy on many NLP tasks including text classification [Dai et al., 2019], language understanding Liu et al. [2019], Wang et al. [2019a,b], machine translation [Lample

and Conneau, 2019], co-reference resolution [Joshi et al., 2020], commonsense inference [Bosselut et al., 2019], and summarisation [Liu et al., 2019] among others. We discuss pre-trained models more in depth in Section 2.1.4.

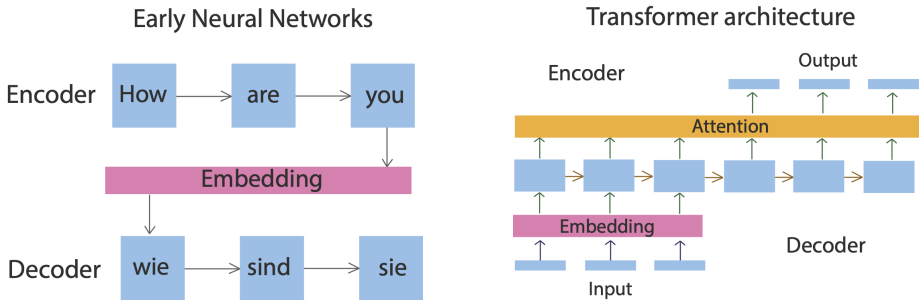

Figure 2.1: A comparison between the sequence-to-sequence architecture of early neural network models such as RNN and transformer architecture.

Pre-trained Word Models

The representation of words has been a long-standing task in NLP [Chiang et al., 2020] where the main underlying principle was based on the idea that the meaning of a word can be understood by the words in its context [Firth, 1957]. However, traditional NLP techniques represent words simply as indices in a vocabulary without a notion of similarity between words. This type of representation provides simplicity and robustness. Such simple models trained on huge amounts of data outperform complex systems trained on less data [Brants et al., 2007, Mikolov et al., 2013a]. An example of such model is the N-gram model used for statistical language modeling and also used for the creation of feature vectors which are the basis for the development of statistical ML models [Brants et al., 2007, Mikolov et al., 2013a]. However, frequency-based word representations are at their limits in many tasks [Brants et al., 2007, Mikolov et al., 2013a] as the amount of domain-specific data for many tasks is limited. Thus, there are situations where simple scaling up of the basic techniques will not result in any significant improvement in model's performance, and there is a need for more advanced techniques [Mikolov et al., 2013a].

With progress of deep ML techniques, it has become possible to train more complex models on much larger unlabelled data sets which often outperform the simple models [Turney and Pantel, 2010]. Probably the most successful concept is to use distributed representations of words [Hinton et al., 1986, Feldman and Ballard, 1982] where distributed representation describes the same data features across multiple scalable and interdependent layers. This principle is utilised by NNs for the creation of low-dimensional word representations learned from text corpora (i.e. word embeddings) [Mikolov et al., 2013a, Pennington et al., 2014, Bojanowski et al., 2017]. The NN-based word representations aim at capturing similarities between words and outperform N-gram models for many NLP tasks [Mikolov et al., 2013a].

Neural word embeddings have been proven to contain useful information about concepts and entities, and provide a generalization boost to many NLP applications [Goldberg, 2017, 2016]. The word representations computed using NNs encode many patterns and linear relationships between words in the vector space, which are demonstrated by analogy [Chiang et al., 2020]. For instance, the result of a vector calculation *vec('king')-vec('man')+vec('woman')* will result in vector representation close to the vector representation for *'queen'* [Mikolov et al., 2013c,a].

Additionally, model pre-training [McCann et al., 2017, Howard and Ruder, 2018, Devlin et al., 2019, Beltagy et al., 2020, Liu et al., 2019] allows models to be trained on unlabelled generic corpora and subsequently be easily adapted to specific tasks. Pre-training on large unlabelled datasets allows vector representations for words that do not appear in the supervised training set. However, the representations for these words might be similar to those of related words that do appear in the training set which allows the model to generalise better on unseen data [Goldberg, 2016]. Earlier pre-trained models build using traditional neural network architectures are called word embeddings (see Section 2.1.4) while language models are build using Transformer implementation principles (see Section 2.1.4).

Recent Word Embedding Models Word Embeddings are representations of words as low-dimensional vectors that capture the semantic relationships between words. Efficient methods for learning high-quality vector representations of words from large amounts of unstructured text data by using NNs are Continuous Bag-of-Words (CBOW) and skip-gram models [Mikolov et al., 2013a,b]. The Skip-gram model learns to predict a target word thanks to a nearby word. On the other hand, the CBOW model predicts the target word according to its context. Unlike most of the previously used NN architectures for learning word vectors, training of the skip-gram model does not involve dense matrix multiplications which makes the training more efficient [Mikolov et al., 2013b]. An example of how the two approaches work is given in Figure 2.2. Popular model using the skip-gram approach for building term representations is Word2Vec [Mikolov et al., 2013b]. Word2Vec [Mikolov et al., 2013b] is a computationally efficient two-layer neural network model for learning term embeddings from raw text. The output of the model is an embedding matrix, where each term (single or multi-token) from the corpus vocabulary is represented as an n-dimensional vector. A limitation of Word2Vec is that it ignores the morphology of words by assigning a distinct vector to each word [Bojanowski et al., 2017]. This limitation is addressed in fastText [Bojanowski et al., 2017] approach where each word is represented as a bag of character n-grams. A vector representation is associated with each character n-gram and words are represented as the sum of these representations. This allows to build vectors for rare words, misspelled words or concatenation of words. Another widely used word embedding model is Glove [Pennington et al., 2014]. It is a log-bilinear model with a weighted least-squares objective. It is a hybrid method that uses machine learning based on statistic matrix which makes it less time efficient than the other approaches. Word embedding models, pre-trained on large corpora of unlabelled data such as news corpora, are widely used in solving NLP problems by fine-tuning them to the specific task.

A limitation to word embedding models is that they use unidirectional approaches for learning word representations and thus they produce a single vector per word despite

the context in which it appears.

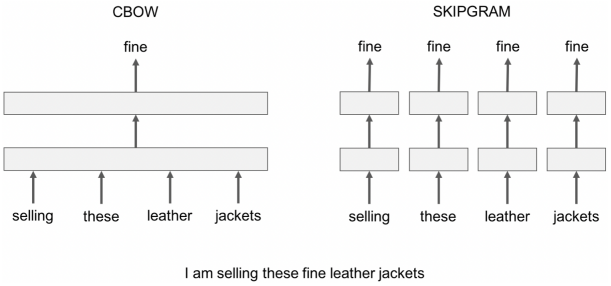

Figure 2.2: Comparison between cbow and skip-gram approaches: Given the sentence 'Selling these fine leather jackets' and the target word 'fine', a skip-gram model tries to predict the target using a random close-by word, like 'leather' or 'these'. The cbow model takes all the words in a surrounding window, such as [selling, these, leather, jackets], and uses the sum of their vectors to predict the target.

Language Models As mentioned in Section 2.1.4 word embeddings, such as Word2vec suffer from the limitation of being context insensitive, i.e., the word is associated with the same representation in all contexts, disregarding the fact that different contexts can trigger various meanings of the word, which might be even semantically unrelated. The more recent Transformer-based contextualised embeddings [Peters et al., 2018, Devlin et al., 2019] address this limitations by computing dynamic representations for words based on the context in which they are used. Further, their scalability allow these models to be efficiently pre-trained on large corpora and then adapted to downstream tasks through fine-tuning [Peters et al., 2019].

One of the first state-of-the-art language models is Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2019] which overcomes the unidirectionality constraint associated with word embeddings by using transformer-based Masked Language Model (MLM) which randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Unlike left-to-right language model pre-training, the MLM

objective enables the representation to incorporate the left and the right context, which allows more context-based representations. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications [Devlin et al., 2019]. BERT model has been pre-trained using BooksCorpus (800M words) Zhu et al. [2015] and English Wikipedia (2,500M words). There are two steps in adapting a language model to a specific task framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabelled data over different pre-training tasks. For fine-tuning, the BERT model is first initialised with the pre-trained parameters, and all of the parameters are fine-tuned using labelled data from the downstream tasks. Each downstream task has separate fine-tuned models even though they are initialised with the same pre-trained parameters (see Figure 2.3). BERT has proved to provide state-of-the-art performance against most standard NLP benchmarks [Wang et al., 2019a,b, Gururangan et al., 2020, Rogers et al., 2020], including text classification.

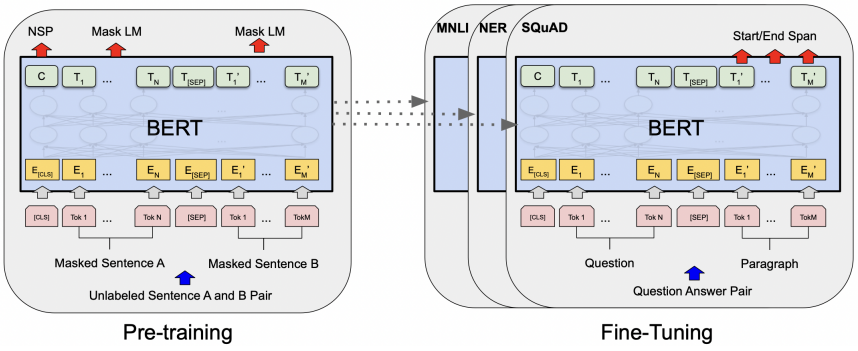

Figure 2.3: BERT model architecture (pre-training and fine-tuning steps [Devlin et al., 2019]: apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token..

A limitation to pre-trained models such as BERT is that they may require fine-tuning

on a large volume of task-specific data to achieve strong performance on the given task. The need for a large dataset of labelled examples for every new task limits the applicability of language models [Brown et al., 2020].

Bigger and more recent language models such as GPT and its recent releases GPT-2 [Radford et al., 2019] and GPT-3 [Brown et al., 2020] address these limitations of earlier language models by introducing a zero-shot learning objective. GPT model is a large transformer-based language model trained on a dataset of 8 million web pages [Radford et al., 2019]. It is a feed-forward generative model which makes it suitable for predicting the next token in a sequence in contrast to BERT architecture where the model is bidirectional (see Figure 2.4). GPT model has been used successfully in text generation tasks such as summarising [Xiao et al., 2020, Kieuvongngam et al., 2020, Alambo et al., 2020, Wang et al., 2019c] and question answering [Liu and Huang, 2019, Baheti et al., 2020, Klein and Nabi, 2019]. GPT-2 is the most recent release of the model (GPT-3 has not been released yet) pre-trained on 40GB dataset.

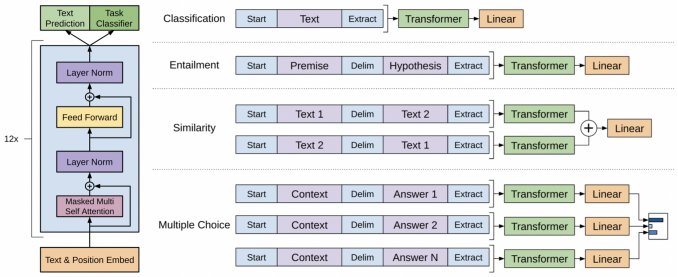

Figure 2.4: GPT architecture [Radford et al., 2018]: Transformer architecture(left) and Input transformations for fine-tuning on different tasks (right), where all structured inputs are converted into token sequences to be processed by the pre-trained model, followed by a linear+softmax layer.

Considering the wide usage of pre-training and fine-tuning techniques for language models for wide range of NLP tasks, we continue the discussion on this topic more in depth in Section 2.2.

2.1.5 Summary

In this section, we have identified three main IE approaches, traditional IE, early neural network models, and transformer models (see Table 2.1 for a comparison). NN architectures, especially pre-trained language models give state-of-the-art performance for many NLP tasks such as question answering, machine translation, reading comprehension, and summarisation. However, adapting pre-trained language models to the task still require large volumes of labelled datasets related to the task. Recent work on creating generative transformer-based models such as GPT, suggests that task-specific architectures are no longer necessary [Radford et al., 2018, Devlin et al., 2019]. However, the full potential of such models for text classification and specialised domains, especially when datasets are scarce has not been fully investigated. Further, traditional IE approaches are still a preferred method for many specialised domains due to their less resource-consuming nature. Despite this, there is a lack of investigation into how NN models and earlier traditional classification approaches compare when used for small and specialised domains. We address this research gap with research question **RQ 2**.

2.2 Adapting Pre-trained Models to Domains and Tasks

Fine-tuning contextualised word embedding models such as BERT [Devlin et al., 2019] on a particular task has become the new standard approach, replacing the more traditional knowledge-based and fully supervised approaches [Sainz and Rigau, 2021]. However, specialised domains, such as the medical domain [Lee et al., 2020], contain large number of domain-specific terminology which are understood mainly by experts within the domain. As a result, NLP models designed for general purpose language understanding often obtain poor performance in domain-specific tasks [Lee et al., 2020, Chakrabarty et al., 2019, Huang et al., 2019]. Thus, a recent research is focused on how to adapt these large but generic models to specific domains and tasks as well as

IE method	**Characteristics**	**Advantages**	**Disadvantages**
Traditional IE	Involve the use of lexical resources, rules, or statistical machine learning algorithms for performing IE tasks	Computationally inexpensive; satisfactory results with less/no training data; successfully used for specialised domains when domain-relevant lexical resources are available	No transfer learning; cannot deal with OOV words; no semantic relations between sequences
Early neural models	IUse a feed-forward approach, which processes the words of text input in a sequential manner with one word followed by the next word	Transfer learning; semantic representations; learn complex relationships	Resource consuming; struggle to capture long term dependencies; unclear benefits for domains with limited unlabelled and labelled data
Transformer models	Represent text in a non-sequential manner where the representation of each word is directly connected to the representation of every other word (use attention mechanism that update one representation as a function of other connected representations)	Transfer learning; contextual representations; easy to fine-tune to different tasks	Resource consuming; unclear benefits for domains with limited unlabelled and labelled data

Table 2.1: A comparison between the three main types of Information Extraction techniques, i.e 'Traditional IE', 'Early neural models', and 'Transformer models'.

investigating to what extend continuous pre-training and fine-tuning are helpful for improving their performance.

We mainly distinguish between two types of adapting language models, i.e., domain-adaptive pre-training (DAPT) and task-adaptive pre-training (TAPT). DAPT refers to pre-training on unlabelled domain data while TAPT refers to pre-training on the unlabelled training set for a given task [Gururangan et al., 2020]. TAPT uses a far smaller pretraining corpus, but one that is much more task-relevant [Gururangan et al., 2020].

Domain-adaptive Pre-training Most of the research on DAPT approaches is aiming to provide language models that fit the terminology of the medical and clinical domain. For instance, Alsentzer et al. [2019] release BERT models for clinical text: one for generic clinical text and another for discharge summaries specifically. They show that using a domain-specific model yields performance improvements on three common clinical NLP tasks as compared to nonspecific embeddings.

Similarly, Lee et al. [2020] create BioBERT by pre-training the generic BERT model further on biomedical text such as PubMed abstracts and PMC articles. BioBERT is fine-tuned and evaluated on three popular biomedical text mining tasks, i.e., NER, relation extraction and question answering. Further, a research by Chakrabarty et al. [2019] fine-tune a language model using Reddit corpus of 5.5 million opinionated claims to improve the task of claim detection. Empirical results show that using the Reddit corpus for language model fine-tuning improves the state-of-the-art performance across four benchmark argumentation datasets.

Additionally, Chalkidis et al. [2020b] investigate different strategies for applying BERT model in the legal domain. These strategies are: using the original BERT out-of-the-box, adapt BERT by additional pre-training on domain-specific corpora, and pre-train BERT from scratch on domain-specific corpora. Results showed that adapting BERT to the specific domain is important for achieving satisfactory results in domain related tasks.

However, adapting language models to the specific domain can be still a very highly data-consuming process unsuitable for domains with sparse collections. These scenarios are not considered in the aforementioned research where authors assume access to large amounts of data.

Task-adaptive Pre-training Research on TAPT [Sun et al., 2019, Logeswaran et al., 2019, Han and Eisenstein, 2019, Chronopoulou et al., 2019, Radford et al., 2018] showed the benefit of continuous pre-training and fine-tuning language models (mainly

investigated BERT) onto the task-specific dataset. For instance, Sun et al. [2019] explore several ways of fine-tuning BERT to enhance its performance specifically on text classification task, including few-shot scenarios. The authors found that with further pre-training BERT performs well in few-shot text classification [Sun et al., 2019] when it has been further pre-trained on the domain dataset. Other TAPT approaches include language modeling as an auxiliary objective to task classifier fine-tuning [Chronopoulou et al., 2019, Heap et al., 2017] or consider the syntactic structure of the input while adapting to task-specific data [Swayamdipta et al., 2019].

A main drawback to the DAPT and TAPT approaches presented above is that they are limited to a single domain or task. Gururangan et al. [2020] provide a more extensive research on DAPT and TAPT approaches by covering four domains and eight classification tasks. The model used for performing analysis is RoBERTa [Liu et al., 2019]. Results showed that a second phase of DAPT leads to performance gains, under both high and low resource settings. Moreover, the authors showed that TAPT improves performance even after domain-adaptive pretraining. Gururangan et al. [2020] also investigate how the performance of continued pre-training may vary with factors like the amount of available labelled task data, or the proximity of the target domain to the original pre-training corpus.

Metaembeddings In order to avoid pre-training and fine-tuning methods as these can be data-consuming and computationally expensive approaches, some research is focused on improving embedding and language models for specific domains and tasks by using metaembeddings. Metaembeddings are build by combining different embedding models, in order to improve their coverage and performance [Yin and Schütze, 2016]. The ensemble approach has two benefits. First, enhancement of the representations — metaembeddings perform better than the individual embedding sets. Second, coverage — metaembeddings cover more words than the individual embedding sets. Specifically, Yin and Schütze [2016] propose a simple method of improving word embeddings for small corpus based on simply averaging corpus-trained and pre-trained word em-

beddings. Despite its simplicity, this approach has proved to outperform more complex algorithms. Similar research to the one presented by [Yin and Schütze, 2016] have tried to improve performance on specific tasks by using several embedding sets simultaneously [Tsuboi, 2014, Turian et al., 2010, Luo et al., 2014]. Additionally, the authors of Li et al. [2018] aim at building 'generalised' classifiers for filtering crisis tweets and perform experiments using different word embedding models. They propose two approaches for building tweet vectors, one based on combining word embedding models and another using sentence encoding methods. Experiments are performed using a Gaussian Naive Bayes(GNB) classifier and SVM.

2.2.1 Summary

A main gap in current research on adapting language models to domains and tasks is the lack of extensive analysis of how generic language models perform for text classification with limited data in comparison to other approaches, such as statistical machine learning and embeddings. Further, there is a lack of research with a specific focus on investigating these models performance for few-shot text classification. Most of current research assume large amounts of unlabelled or labelled data or even both. These gaps in recent studies on language models and transfer learning make it hard to identify when and how these pre-trained models can be applied to domains with limited data, especially for text classification. We aim to address this research gap by answering **RQ 3**.

2.3 Text Classification

2.3.1 Definition and Applications

Text classification is a fundamental research area in NLP [Lyu et al., 2020] as it is one of the most important methods to organise and thus help use the large amounts

of information that exists in unstructured textual format [Altinel and Ganiz, 2018]. Text classification, also referred to as text categorisation is using supervised machine learning techniques in order to assign one or more class labels or categories from a predefined set of labels or categories to a given text, according to its content [Deng et al., 2019, Kong et al., 2019, Zhong and Enke, 2019]. Supervised machine learning techniques are very often applied in web-based information retrieval systems for classifying web pages and news, recommend systems for suggesting items to users based on the description of an item and a profile of the user's interests [Aggarwal, 2016], and for information filtering such as spam emails filtering [Deng et al., 2019, Aggarwal and Zhai, 2012].

Text classification is also highly valuable in more specialised domains such as medicine, social sciences, healthcare, psychology, and law [Kowsari et al., 2019]. For instance, in the social science domain, text classification and document categorisation have increasingly been applied to understanding human behavior [Nobles et al., 2018, Ofoghi and Verspoor, 2017]. In particular, recent research in human behavior have focused on mining language contained in informal notes and text data sets, including short message service (SMS), clinical notes, social media, etc [Nobles et al., 2018]. In the domain of law, there are large volumes of legal documents generated by government institutions that require automatic approaches for structuring them to support lawyers in their work. The categorization of these documents is the main challenge for the lawyer community [Turtle, 1995].

Applying text classification to specialised domains is highly challenging and not widely researched area. However, the high need for establishing supervised approaches for organising such texts is the main motivation for this thesis.

The main steps of the text classification process involve Feature Extraction, Feature Integration, and using a classification algorithm to build a predictive model for labeling unseen text instances (see Figure 2.5). Throughout the thesis, we perform various experiments involving all three main steps of the text classification process. We look

at how different information extraction and integration methods affect the classifier's performance. Therefore, in the following sections, we go through each step explaining main existing approaches.

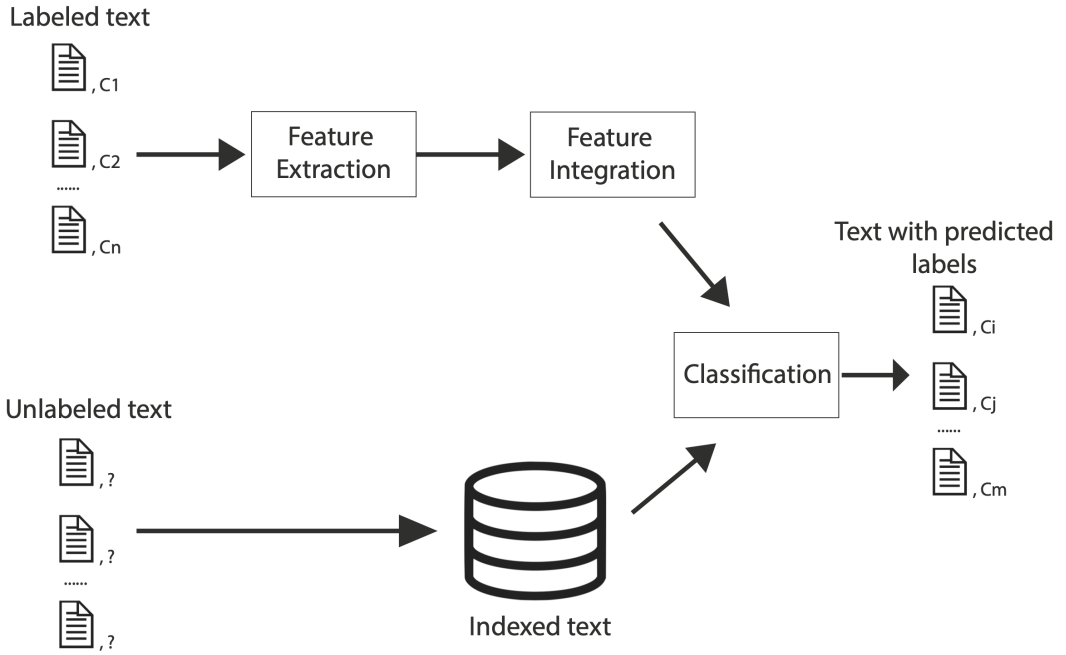

Figure 2.5: Text Classification Process Overview

2.3.2 Feature Extraction

Unstructured texts must be converted into a structured feature space when using mathematical modeling as part of a classifier [Kowsari et al., 2019]. First, the data needs to be cleaned to omit unnecessary characters and words. After the data has been cleaned, formal feature extraction methods can be applied [Kowsari et al., 2019]. The common techniques of feature extractions can be divided into two groups: count-based techniques such as word frequencies and TF-IDF [Salton and Buckley, 1988] and neural network-based word models such as word embeddings (Word2Vec, fastText, Glove) and language models such as BERT. As explained in Section 2.1.4, simple frequency-based methods are easy to compute but do not capture position in text. In contrast, NN

models are resource-consuming to compute but they can capture positions of words and contextual meaning.

2.3.3 Feature Integration

Feature integration often involves dimensionality reduction. Feature integration is defined as the step of combining word feature vectors into a single feature vector used by the classification algorithm. A simple feature integration technique is bag-of-words (BOW) where the order of words is not preserved. The BOW model is used as the standard representation of text input for many statistical classification models such as SVM [Cortes and Vapnik, 1995] and Naive Bayes classifiers [McCallum and Nigam, 1998]. Statistical classifiers have been widely studied for many text classification problems [Sebastiani, 2002], such as document classification or sentiment analysis, due to their efficiency, robustness and interpretability, and the BOW text representation can capture sufficient information for statistical classifiers to make highly accurate predictions [Dumais et al., 1998]. However, in settings where there is a large vocabulary, a high number of classes (e.g., complex ontologies) and short text (e.g., fragments of text, single sentences or document titles) the BOW representation contains extremely sparse data which reduces the accuracy of the linear classification models [Wang and Manning, 2012]. A particular problem with classification of short texts is rare words, or as an extreme, words that do not occur at all in the text used to train models, but do occur in test data [Heap et al., 2017].

2.3.4 Classification Algorithms

We outline some of the main algorithms used or considered through this thesis, summarised by [Aggarwal and Zhai, 2012, Colace et al., 2014, Sebastiani, 2002].

Decision Trees Decision tree [Quinlan, 1986] is one of the earlier classification algorithms used for classifying text. It is based on a hierarchical decomposition of the training data space where a condition on each attribute value is used to divide the data space into class partitions. Decision trees are suitable for handling categorical features. However they are extremely sensitive to small perturbations in data and are exposed to problems with out-of-sample predictions and overfitting [Kowsari et al., 2019].

Logistic Regression This algorithm does not require pre-processing or tuning of input features. However, it cannot solve non-linear problems and requires data points to be independent [Kowsari et al., 2019].

SVM SVM classifier [Cortes and Vapnik, 1995] partitions the data space into classes using either a linear or non-linear function. SVM is a robust classifier against overfitting, however its performance is dependent on finding an optimal boundaries between the different classes. Further, it lacks transparency in results caused by high number of dimensions (especially for text data) [Kowsari et al., 2019].

Bayesian Classifiers Bayesian classifiers are probabilistic classifiers. In particular, Bayesian probability is applied towards prediction of the value of the dependent variable, yet without considering any relation- ships or weightings between the independent variables [McCallum and Nigam, 1998]. This types of classifiers work very well with text data and provides explanations on most significant features for making predictions. However, it is limited by data scarcity for which any possible value in feature space, a likelihood value must be estimated by a frequency [Kowsari et al., 2019].

Neural Network Classifiers These classifiers use based on NN architectures, introduced in Section 2.1.2. More recent NN classifiers are based on the use of pre-trained language models which are adapted, i.e., fine-tuned to the classification task, as described in Section 2.2. These classifiers are related to SVM classifiers, because they

are both in the category of discriminative classifiers. A main advantage of using NNs for classification is the ability to model complex, non-linear relations between data features and the ability to leverage a generic language knowledge in the form of pre-trained models. However, as already mentioned earlier, neural network models require large amounts of data and are computationally inefficient [Kowsari et al., 2019]

2.3.5 Data Scarcity in Text Classification

A main problem with existing supervised classification approaches is that they need a significant amount of training data to achieve high results [Christopher et al., 2008, Sebastiani, 2002, Lewis et al., 2004, Lyu et al., 2020, Türker, 2019, Li and Yang, 2018, Cawley and Talbot, 2010, Colace et al., 2014]. However, manual labeling of data is a time-consuming and costly process [Türker et al., 2019, Zhang and Wu, 2015, Shams, 2014, Kumar et al., 2020], especially when the text to be labelled is from a highly-specialised domain where only scarce domain experts can perform the labelling task [Türker et al., 2019, Ali, 2019, Marivate and Sefara, 2020]. For instance, if the text to be labelled is of a specialised domain, crowd-sourcing based labeling approaches do not work successfully and only expensive domain experts are able to fulfill the manual labeling task [Fernandes de Araújo et al., 2020, Zhang et al., 2020b]. Therefore, a wide area of research in text classification is focused on overcoming the data scarcity and class imbalance problems associated with it.

NN-based classifiers have achieved great success in text classification [Gururangan et al., 2020, Rogers et al., 2020] even for more specialised domains such as the medical domain. For instance, Song et al. [2020] present an approach for identifying suicidal behaviour from the free text part of electronic health records using a combination of word embeddings, LSTM, and CNN models.

However, as mentioned in Section 2.1.2, NN models usually require large amounts of manually annotated data, which greatly limits the practicality and scalability of such

models [Lyu et al., 2020, Yang et al., 2020, Strubell et al., 2019, Sainz and Rigau, 2021] for domains with sparse amount of data. Recent Transformer models such as GPT-3 [Brown et al., 2020] shows that when increasing the size of the model, the capacity to solve different tasks with just a few positive examples also increases [Sainz and Rigau, 2021] through the use of fine-tuning techniques and transfer learning methods.

As the thesis is mainly concerned with data scarcity problem in text classification, in the following sections we will review related research on this topic as follows: In Section 2.3.6), we look at recent research on developing classification methods for low resource settings such as zero-shot classifiers and few-shot classifiers. In Section 2.3.7 we reflect on techniques used to enhance classifiers performance by enriching feature vectors using external semantic knowledge. Section 2.3.8 present recent techniques on using data augmentation strategies.

2.3.6 Low Resource Text Classification

Zero-shot Classification

Zero-shot classification or zero-shot learning (ZSL) also referred to as dataless classification Mylonas et al. [2020] is a method which do not require labelled data as training instances Li and Yang [2018]. Instead, ZSL aims to classify documents of classes which are absent from the learning stage [Zhang et al., 2019]. Zero-shot classification is expected to exploit supportive semantic knowledge such as class descriptions, relations among classes, and external domain knowledge, in order to infer features of unseen classes [Zhang et al., 2019].

ZSL is mainly used in situations where a classification framework is susceptible to frequent changes [Zhang et al., 2019, Ye et al., 2020, Chalkidis et al., 2020a] such as insertion, deletion or change of some of the classes. It is also used when a supervised classification had been performed for a related task [Ye et al., 2020] where an existing classifier can be adjusted to changes of the class framework or a similar task.

Approaches are mainly investigating relations between seen and unseen classes and transferability of approaches.

Another group of ZSL methods, presented in [Chang et al., 2008, Song and Roth, 2014, Türker et al., 2019, Türker, 2019] is motivated by scenarios where there is no existing supervised classifier for a similar task. In such scenarios, approaches heavily rely on a semantic similarity between a given text and a set of predefined categories to determine which category the given document belongs to. More specifically, documents and categories are represented in a common semantic space based on the words contained in the documents and category labels which allow us to calculate a semantic similarity between documents and categories. A downfall of these approaches is that they rely on the existence of the classification categories in a publicly available knowledge base (e.g. Wikipedia), which might not be the case for more domain-specific corpora.

Overall, zero-shot approaches rely on either a similarity between unseen classes and seen classes or some semantic similarity between knowledge base terms and unseen classes. We do not go in depth reviewing approaches as we feel the motivation for these is outside the scope of our research. We focus on highly-domain specific texts where there is some amount of labelled data and there are no similar classification tasks which can be adapted to the given task.

Few-shot Classification

Recently, there has been an increased motivation to tackle the problem of data scarcity for text classification using Few-shot Learning (FSL) [Gupta et al., 2020, Wang et al., 2020, Miller et al., 2000, Fei-Fei et al., 2006, Lake et al., 2015]. This method uses prior knowledge to generalise to new tasks containing only a few labelled instances [Wang et al., 2020, Miller et al., 2000, Fei-Fei et al., 2006, Wang et al., 2020, Bailey and Chopra, 2018, Gupta et al., 2020]. A popular FSL scenario is where examples with supervised information are hard or impossible to acquire [Altae-Tran et al., 2017], similarly to the safeguarding domain.

Although FSL has been explored more in the domain of computer vision, recent work in developing FSL methods for NLP tasks have emerged [Gupta et al., 2020]. Additionally, there have been a widespread growth of interest in transfer learning for NLP, with transformer-based models achieving strong results on a variety of benchmark problems. These models can be fine-tuned to new tasks with a small number of training examples, suggesting that their generalisation capabilities may also be applicable to the few-shot learning setting. Therefore, few-shot learning often goes hand-in-hand with transfer learning [Bailey and Chopra, 2018].

A widely explored area for few-shot text classification is based on the use of prototypical networks which learn a metric space in which classification can be performed by computing distances to prototype representations of each class [Vinyals et al., 2016, Snell et al., 2017, Satorras and Estrach, 2018, Sung et al., 2018, Yu et al., 2018a, Schick and Schütze, 2020]. For instance, Bailey and Chopra [2018] propose a human-in-the-loop approach where a one or two manually labelled documents are used as prototypes (i.e, best representatives) of the given classes. This approach represents documents using pre-trained word embeddings and then uses Latent Dirichlet Allocation (LDA) in order to identify most likely representative documents per category. The selected documents are presented to the user who must manually classify some of the documents for each category. The obtained documents are used to assign a representative vector to each category. The remaining documents are compared against each category using cosine similarity and each one is assigned the category whose vector it is closest to [Bailey and Chopra, 2018]. A more recent technique based on prototypical network [Schick and Schütze, 2020] is based on identifying words rather than documents that can serve as representatives for labels given small amounts of training data. Further, authors use an automatic approach for selecting class representative words based on converting textual inputs to cloze questions that contain some form of task description, process them with a pre-trained language model and map the predicted words to labels.

The prototypical-based approaches, described by Bailey and Chopra [2018] and Schick and Schütze [2020] heavily rely on the embedding models to fit well the domain at hand. However, they use pre-trained embedding and generic datasets which makes the efficiency of these approaches in real applications unclear. Further, relying on a similarity score between a few instances and unlabelled data might not work well where data is with a diverse vocabulary or when data to be classified consists of short sentences rather than documents.

Many related machine learning approaches have been proposed for use in FSL, such as meta-learning [Ravi and Larochelle, 2017, Finn et al., 2017, Mishra et al., 2018, Geng et al., 2019, Bansal et al., 2020, Deng et al., 2020, Santoro et al., 2016] which aim to modify the optimization strategy to provide a model that can rapidly adapt to related tasks. For instance, Yu et al. [2018b] proposed an adaptive metric learning model, which can automatically determine the best weighted combination of a set of metrics obtained from a meta-learning process for a newly arrived few-shot text classification task. Gao et al. [2019] proposed prototypical network by adopting hierarchical attention mechanism, which is applied in feature level, word level and instance level to enhance the expressive ability of semantic space. While Geng et al. [2019] applied the dynamic routing algorithm in meta-learning and proposed an induction network, which achieves a better generalization ability on different few-shot text classification tasks.

Many approaches explored GNNs for few-shot learning [Gori et al., 2005, Scarselli et al., 2008, Bruna et al., 2014, Henaff et al., 2015, Defferrard et al., 2016, Satorras and Estrach, 2018, Kim et al., 2019, Gidaris and Komodakis, 2019]. However, most of these methods use static word embedding models and focus more on the semantic features of the texts itself, ignoring the potential relationships between texts Lyu et al. [2020]. Further, they use RNN or CNN-based neural networks. The approach proposed by Lyu et al. [2020] build on such approaches by creating text embeddings using BERT, edge-labeling graph neural network component and prototypical network com-

ponent. Their method is Evaluated on Amazon Reviews and Relation classification datasets [Lyu et al., 2020].

Another group of FSL-related work is based on using transfer learning approaches and recent Transformer-based models. For instance, Raffel et al. [2020] demonstrate that pre-training on in-domain unlabelled data can improve performance on downstream tasks, suggesting that the initial training data available to few-shot learners could also be used to improve generalization of pretrained models to few-shot classes. Further, a research by Gupta et al. [2020] study a transfer-learning approach applied specifically to few-shot classification. The authors use a simple BERT-based classification scheme that first fine-tunes a pretrained model on the full rating classification dataset, and then further fine-tunes on only the held-out few-shot classes. This approach achieves comparable performance to state-of-the-art techniques, suggesting that pretrained models can extend their generalization capabilities to few-shot settings. Surprisingly, however, Gupta et al. [2020] found similar performance in zero-shot settings for the Amazon review sentiment classification, implying that few-shot categories are not sufficiently distinct from the other categories, and consequently motivating the need for new datasets to support future research in this area.

Summary

Most of the approaches described above are using transfer learning techniques, similar to those discussed in Section 2.2 for addressing the lack of labelled data for classification tasks. Similar to limitations of DAPT and TAPT approaches presented in Section 2.2, the research is limited in scale and there is a lack of comparison between traditional and more recent state-of-the-art methods. Further, there is lack of extensive analysis on how these FSL methods perform for different few-shot classification scenarios, especially with less than 20 instances per label.We address these gaps in literature review with **RQ 3**.

2.3.7 Enriching Feature Vectors using Lexical Resources

The use of publicly available lexical resources for enriching features vectors to improve classification is a widely used method due to its simplicity as it doesn't require the creation of domain-based knowledge graphs and it has proven to improve classification performance.

The approaches described in [Faruqui et al., 2015, Mrkšić et al., 2017] exploit semantic relations between words from lexical resources, such as WordNet, in order to tune word vector spaces so semantically similar words have similar vectors. In both papers they use WordNet to enrich the word vectors, and also perform analysis with other resources such as PPDB or BabelNet. However, WordNet-based augmentation lead to better results.

Choi et al. [2017] address data insufficiency and interpretation of deep learning models for the prediction of rarely observed diseases. For these purposes, they use a neural network with graph-based attention model that exploits ancestors extracted from the OWL-SKOS representations of ICD Disease, Clinical Classifications Software (CCS) and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). Other research [Min et al., 2017, Salguero et al., 2018] use publicly available bio-ontologies for enhancing performance of machine learning approaches for predicting activities of daily living for cancer patients. The ontology-guided ML method was more accurate at predicting ADL performance levels than methods without ontologies in both cases.

Gazzotti et al. [2019] use linguistic resources for enhancing classification performance for predicting hospitalisation. They use specialised ontologies, DBPedia and Wikidata to enrich the features extracted from electronic medical records used by the machine learning algorithms to predict hospitalisation. Using knowledge bases improve the classification task. Further, the same authors 2020, extend on this research by proposing a semi-supervised method for filtering relevant domain knowledge from a general knowledge source. Their main goal is to provide a method to solve the problem of

retrieving relevant knowledge in the medical domain from general knowledge source. The authors use SVC and Random Forest (RF) classifier.

All the approaches described above involve the existence of a knowledge resource which fits the needs of the domain or otherwise these approaches require building lexical knowledge base. However, building large and rich lexical knowledge bases is a very costly effort which involves large research groups for long periods of development [Sainz and Rigau, 2021]. Further, a challenge in using large lexical resources, given the amount of general information available on DBpedia, is to filter the knowledge which is specific to the given domain [Gazzotti et al., 2020].

Summary

Enriching feature vectors using lexical resources for improving text classification has proven effective especially in the medical domain. The high availability and easy access of such resources makes them a preferable approach for enhancing classification performance. However, their generic nature might make them unsuitable for more specialised texts. Despite this, the high effectiveness of such approaches to medical-related data such as the medical domain and their less-resource consuming nature compared to neural network models motivated an investigation into whether such an approach can perform well for more specialised domain such as the safeguarding domain. We address this research gap with **RQ 1**.

2.3.8 Data Augmentation Strategies for Classification

Definition

The collection of labelled data can be specifically challenging for tasks such as prediction of rare diseases and fraud detection where data availability is highly dependent of the occurrence of infrequent events for the the acquisition of a sufficient amount

of labelled data might be impossible. In such scenarios, using automated methods for generating additional training data becomes increasingly important [Kumar et al., 2020, Marivate and Sefara, 2020].

Data Augmentation (DA) is a widely used method for tackling data scarcity and class imbalance problems for classification tasks [Wong et al., 2016, Anaby-Tavor et al., 2020, Kumar et al., 2020, Papanikolaou and Pierleoni, 2019] by synthesizing new data from existing training data with the objective of improving the performance of the downstream model [Anaby-Tavor et al., 2020]. It is a low-cost method for obtaining additional labelled data [Liu et al., 2020, Xu et al., 2020] and also can help avoid class overfitting [Anaby-Tavor et al., 2020]. DA techniques are well established in the domains of computer vision and speech recognition where it is easy to generate new data by simple image transformations such as cropping, padding, flipping, and shifting along time and space dimensions. Such transformations are class preserving when applied on image data [Anaby-Tavor et al., 2020, Giridhara et al., 2019, Krizhevsky et al., 2017, Cui et al., 2015, Ko et al., 2015, Szegedy et al., 2015]. However, such simple techniques cannot be directly applied to text as they can lead to syntactic and semantic distortions and thus the label of original text might not be preserved [Giridhara et al., 2019, Anaby-Tavor et al., 2020]. This makes the development of DA methods for text classification a challenging task. Further, it is very difficult to obtain universal rules for transformations which assure the quality of the produced data and are easy to apply automatically in various domains [Wei and Zou, 2019, Kobayashi, 2018].

Further, human rephrasing is too expensive and unrealistic, whereas machine paraphrasing currently has its limitations. For instance, it only works on specific tasks [Wang and Yang, 2015, Hou et al., 2018] and a specific paraphrase corpus may be required [Fader et al., 2013, Qiu et al., 2020].

Many publications refer to DA methods applied to text as text augmentation (TA) [Sharifirad et al., 2018, Marivate and Sefara, 2020, Liu et al., 2020, Liu et al., 2020] while others keep referring to them as DA methods [Anaby-Tavor et al., 2020, Giridhara et al.,

2019, Krizhevsky et al., 2017, Cui et al., 2015, Ko et al., 2015, Szegedy et al., 2015]. For the rest of the thesis we will use the term Data Augmentation (DA) as it is more widely accepted. DA approaches have been applied to various tasks such as identifying sexist Tweets [Sharifirad et al., 2018], classify emergency-related Tweets [Malandrakis et al., 2019], fake news detection [Krishnan and Chen, 2018], and hate speech discovery [Rizos et al., 2019]. Further, DA has been used for enhancing readability of web documents [Chung et al., 2013], relation extraction [Papanikolaou and Pierleoni, 2019, Kumar et al., 2020], and classification of complaint reports [Sano et al., 2015], and tackle class imbalance problems for extreme multi-label classification tasks [Zhang et al., 2020a], and augment domain-specific datasets in order to improve performance in various domain-specific classification tasks Amin-Nejad et al. [2020].

The existing approaches in DA studies can be split into three main groups: word-replacement based strategies (see Section 2.3.8), sentence-replacement based strategies (see Section 2.3.8), and text generation-based strategies (see Section 2.3.8).

Word-replacement Methods

Word-level transformations can be leveraged to produce new sentences while preserving the semantic features of the original texts to a certain extent Xu et al. [2020]. Word replacement-based (WR)-based DA approaches make local changes only within a given sentence, primarily by synonym replacement of a word or multiple words, deleting words or swapping words order [Anaby-Tavor et al., 2020].

A popular word-replacement method using knowledge bases such as WordNet [Miller, 1998] is Easy Data Augmentation techniques (EDA) [Wei and Zou, 2019]. The method consists of randomly choosing one out of four word replacement techniques for a given sentence: synonym replacement, random synonym insertion, and random swap of words within a sentence. EDA method has been tested on five benchmark datasets and it lead to improvements for few-shot text classification.

A problem with such WR strategies is that words having exactly or nearly the same meanings are very few and thus synonym-based augmentation can be applied to only a small percentage of the vocabulary [Kobayashi, 2018]. Further, using a generic knowledge base such as WordNet for synonym replacement might not be applicable for more domain-specific content.

In order to overcome these problems, related research is using language models coupled with label preserving techniques [Wu et al., 2019] or uses a wider range of substitute words by using words predicted by language model according to context [Kobayashi, 2018]. Specifically, the authors of [Wu et al., 2019] use conditional BERT with an extra label-conditional constraint to the MLM. Thus, conditional BERT can be applied to enhance contextual augmentation.

However, WR-based methods still struggle with label preservation [Kumar et al., 2020, Giridhara et al., 2019, Anaby-Tavor et al., 2020]. For example, using a word swapping technique for a sentiment classification task for the sentence: *'a small impact with a big movie'* can lead to *'a small movie with a big impact'*. Using such augmented data for training, with the original input sentence's label (i.e. negative sentiment in this example) would negatively impact the performance of the resulting model [Kumar et al., 2020]. Further, methods that make only local changes to given instances produce sentences with a structure similar to the original ones and thus lead to low variability of training instances in the corpus [Anaby-Tavor et al., 2020].

Sentence-replacement Methods

Sentence replacement-based (SR) methods are based on back-translation strategies where a given sentence is translated to a language and then back to the original language in order to change the syntax but not the meaning of the sentence [Sennrich et al., 2015, Fadaee et al., 2017]. For instance, an original input English sentence is translated to German and then back to English in order to create additional training data. These methods rely on neural machine translation for achieving back translation [Yu et al.,

2018a, Sennrich et al., 2015]. These approaches are not widely used for DA since they do not provide high diversity into the corpus vocabulary. Further, there is a limit to the number of additional instances that can be generated per a given text.

Text Generation Methods

The application of NN to text generation (TG) has achieved great success in many text generation tasks [Bowman et al., 2016, Shen et al., 2018, Zhou et al., 2017, Du et al., 2017, Zhao et al., 2018]. Further, TG methods have the potential to address the issues with the WR and SR strategies for DA by generating completely new instances from the given original samples. This can help diversify the vocabulary of generated data and help boost the performance of DA approaches for classification tasks.

However, applying generative methods as DA strategies is a relatively new research field [Xu et al., 2020, Amin-Nejad et al., 2020]. Some existing approaches based on generative models include using variational autoencoding [Kingma and Welling, 2013], round-trip translation [Yu et al., 2018a], and methods based on generative adversarial networks (GANs) [dos Santos Tanaka and Aranha, 2019]. A research on the combination of GANs [Mirza and Osindero, 2014] and variational autoencoding [Kingma and Welling, 2013] have proved to give satisfactory results [Su et al., 2020] for DA tasks. However, GAN-based models have excelled primarily in image generation rather than in language tasks [Xu et al., 2020].

Xu et al. [2020] perform an investigation of some data augmentation approaches, including simple resampling, word-level transformations, and neural text generation. Among the text generation methods they explore standard Seq2Seq neural generation as well as variational autoencoding-based models that inject additional variation with stochastic latent variables for data augmentation. Conclusions from this study are that the effectiveness of different data augmentation schemes depends on the nature of the dataset under consideration. Further, authors stress that approaches involving GAN

and VAE models are extremely unstable and the model requires very careful tuning to find a balance between diversity and quality [Xu et al., 2020].

Recent text generative models such as GPT-2 [Radford et al., 2019] has been applied successfully in many tasks requiring text generation such as question answering, summarisation, and corpus augmenting [Liu, 2018, Melamud and Shivade, 2019, Gong et al., 2020].

The state-of-the-art performance of GPT-2 on text generation tasks and its objective to fit scenarios with few-shot and zero-shot NLP tasks makes it a preferable method for performing DA in literature.

Many research on TG- based DAT using GPT are focusing on designing label-preserving strategies for the generated additional data [Anaby-Tavor et al., 2020, Wang and Lillis, 2019, Kumar et al., 2020]. For instance, the most widely accepted approach for label preservation is based on prepending the class labels to text sequences during fine-tuning of the Transformer-based model [Wang and Lillis, 2019, Zhang et al., 2020a, Kumar et al., 2020]. Further, Anaby-Tavor et al. [2020] present a language model-based data augmentation (LAMBADA) where additional training data is generated using GPT-2 and then the new instances are filtered using a classifier trained on the original data in order to re-label the sequences and select only those with high confidence score. LAMBADA approach have been evaluated against three baseline methods, EDA, CVAE, and BERT for three datasets. Zhang et al. [2020a] focus on DA for the extreme multi-label classification (XMC) problem where they compare GPT-2-based approach against EDA Wei and Zou [2019]. The authors group examples pairs with the same label sets, then fine-tune the pre-trained GPT-2 to generate label-invariant sequences. In contrast to other work using GPT-based DA, Zhang et al. [2020a] perform analysis for 1%, 5%, 50%, and 100% of the original data. Results showed that when training data is very limited, both rule-based augmentation and GPT-based approach work better than base models. When training data is rich, GPT-based approach still improves over baseline while rule-based systems start to hurt precisions. There-

fore, authors recommend GPT-based DA since it improves more consistently against different training sizes. Other research that has used GPT successfully in DA tasks include [Wang and Lillis, 2019] where authors upsample classes with only a few training instances for improving classification of crisis-related tweets. Further, Kumar et al. [2020] study three types of Transformer-based pre-trained models for conditional data augmentation, such as seq2seq model BART [Wu et al., 2019], auto-encoder model BERT [Devlin et al., 2019] and auto-regressive model GPT-2 [Radford et al., 2019]. Results showed that in a low resource settings for three classification tasks all three DA methods are effective, however BART outperform the other two TG strategies for high-resource settings. Lastly, Amin-Nejad et al. [2020] used GPT-2 to augment clinical texts related to patient records in order for these datasets to be used in downstream classification tasks and where GPT-2 outperformed SOTA vanilla architecture [Lakew et al., 2018] for two classifiers.

The above mentioned studies on TG-DA methods focus primarily on comparison between different TG methods and the implementation of label-preservation techniques for the generated synthetic data samples. However, an important problem with text generation techniques, ignored in the above research, is the possibility of generating noise which decreases the performance of classification models rather than improving it [Yang et al., 2020]. Further, a randomly sampled synthetic dataset may contain examples that are similar to one another along with low-quality generations [Holtzman et al., 2019]. All these show the need for creating methods that help generate higher quality and more diverse artificial training data.

A recent research by [Yang et al., 2020] presents a generative data augmentation method for commonsense reasoning, called G-DAUG. The proposed approach generates synthetic examples using pretrained language models, and selects the most informative and diverse set of examples for data augmentation. In order to ensure that the most informative examples are used for augmentation, the authors use data selection methods based on influence functions, presented in [Koh and Liang, 2017] and a heuristic

to maximize the diversity of the generated data pool. Classification is performed using RoBERTa and generation is done using GPT-2. Similarly to Anaby-Tavor et al. [2020], they use a classifier trained on the original dataset to re-label generated instances. Specifically, there is a lack of research in methods for selecting seed samples which are used to generate artificial data. We believe that using high quality class representative instances in first place to generate artificial data will lead to producing higher quality training dataset on a potentially lower cost in comparison to approach which applies selection methods on the already generated data. However, we believe that devising strategies which help selection of class representative samples from the original data in the first place can already lead to important improvements and has an important efficiency advantage, as it prevents an unnecessary waste of resources and time of generating unused generated documents, especially considering how resource expensive generative language models are [Strubell et al., 2019, Schwartz et al., 2019].

Summary

The success of text generation models such as GPT and consequent releases, for various tasks, lead to an increase research into data augmentation methods based on text generation. These methods are considered superior to word- based and sentence-based replacement DA methods as they introduce more diversity and less grammar distortions to the generated additional data. Further, data augmentation methods have been widely used for improving text classification for small corpora which is the research area we focus on. However, the problem of quality of generated data and applicability of methods for wider range of domains have not been addressed in literature. Further, most of the approaches do not analyse how methods perform for few-shot scenarios where there is only a couple of training instances available (less than 20). Our research into these problems is reflected in **RQ 4**.

2.4 Conclusions

The motivation for the research questions declared in the previous chapter lies in the need for developing methodologies which help improve text classifiers performance for small and specialised corpora, the latter of which is the problem area identified over the previous sections of this thesis.

Much of the most relevant research has adopted transfer learning and fine-tuning as main techniques for coping with low resource classification. However, most of the approaches are based on generic datasets, assume large collections of unlabelled domain data, and have limited coverage of analysis, classification tasks, and datasets. Therefore, the applicability of such methods to small collections of specialised domains remains unclear and thus is the main research topic in the thesis. Additionally, most research on FSL assume the presence of large amounts of unlabelled data or similarity between domains where one of the domains is associated with large amounts of training data which makes the adaption of the models to the new domain and task easier.

On the other hand, many authors are using lexical resources and traditional count-based classifiers for reaching satisfactory performance in more specialised domains, such as the medical domain. However, such approaches are highly depended on whether lexical resources correspond to the needs of the given domain and task.

Further, with the realise of recent state-of-the-art generative language models which have been created with the objective to work in zero-shot settings, new text generation-based data augmentation methods have emerged for supplementing the original training dataset with additional artificial data. Literature review on using such approaches show their potential in outperforming well-established DA methods based on WR strategies. Further, the progress towards the creation of more contextually-aware language models for various IE tasks shows that there might not be a need for creating lexical resources for the specific domains as this can be time-consuming and a domain-

dependent approach.

Despite the benefits of using data augmentation approaches based on text generation techniques, many questions still remain open. For instance, how can the noise in generated artificial data be reduced, how much data can be generated before affecting classification performance negatively, and are such approaches applicable to specialised domains with limited amounts of unlabelled data.

The research gaps summarised above provided a motivation for extensive research into suitability of existing classification approaches for small and specialised texts (addressed by **RQ 1** and **RQ 2**) as well as the importance of adapting pre-trained models to domain- and task-specific data for few-shot classification (addressed by **RQ 3**). This gives the basis for the development of methodology for improving classification methods for few-shot settings using text generation-based data augmentation techniques. The methodology, presented in this thesis, aims to improve quality of generated data by using strategies for selecting class representative samples from the original dataset used to produce additional training instances. We perform extensive analysis, considering multiple strategies, datasets, and few-shot classification settings. Further, we analyse how different approaches of fine-tuning text generation models affect the quality of generated data and consequently the classification performance. The developed methodology aims at addressing the research gaps related to improving quality of generated data for data augmentation methods with a focus on specialised domains (addressed by **RQ4**).

In the following chapter, we give a detailed explanation of the case study of the safeguarding domain and key definitions giving better understanding of the domain. Further, we present early and exploratory work in using well-established NLP tools for IE as well as investigation into applicability of using publicly available lexical resources for the safeguarding domain. These early analysis help identify challenges for extracting knowledge from the specialised documents as well as helping to understand the decisions made in later chapters of the thesis.

Chapter 3

Case Study and Exploratory Work: Traditional Information Extraction

In this chapter, we explain our motivating scenario in more detail and present exploratory work on using traditional IE techniques for extracting knowledge from the collection of safeguarding reports. As highlighted in Section 1.1, we have been working with the CSRI at Cardiff University on the WSR project. The aim of this project is to provide a repository for housing safeguarding reports as well as provide automated tools for predicting expert generated themes within the documents. The repository has been created to support the faster and easier searchability and readability through the growing collection for researchers and practitioners in the safeguarding domain and thus facilitate faster and more accurate decision making and better resource allocation. In early attempts to improve the searchability of the documents and identify important topics of discussion, we focused our attention on extracting entities with the potential to provide indexing as well as use means for identifying important parts of the documents which require more attention by the readers. Further, the existence of predefined thematic framework for annotating the documents required the need for using supervised approaches for labeling new documents with the themes created by experts.

Our initial work focused on using traditional IE techniques and tools for NER and sentiment analysis for identifying parts of the reports that might be with higher importance. This early analyses helped identify main challenges for analysing the collection and also establish next steps for the creation of classification approaches for predicting

the themes.

Further, we perform experiments with statistical classifiers which provide a strong baseline for many predictive models [Joachims, 1998, McCallum et al., 1998, Fan et al., 2008]. They are even known to give higher performance than neural network-based techniques for some domains and tasks [Sahlgren and Lenci, 2016, Roli et al., 1997]. Additionally, as already mentioned in Section 2.3, many domain-specific tasks are using lexical resources such as databases, ontologies, and taxonomies combined with statistical classifiers in order to boost performance of text classification tasks. The simplicity and less-resource consuming nature of this approach makes it a preferred method for domain-related scenarios with limited training data. A drawback of this method is that it is highly dependent on the existence of lexical resources that fit the needs of the domain and task. Therefore, we investigate whether existing lexical resources can be applied to the safeguarding domain for improving classification and whether there are any existing knowledge graphs related to safeguarding topics. This research addresses question **RQ 1: Can publicly available lexical resources be used to support supervised learning for specialised domains?** from the research questions presented in Section 1.2. More specifically, contributions include a study into the applicability of traditional IE approaches such as NER libraries and the use of lexical resources for augmenting classification using non-neural based classifiers. From this research, an initial methodology of next steps is built.

The structure of this chapter is as follows. Section 3.1 explains our motivation and aim for analysing the safeguarding reports and also describes the dataset. Section 3.2 outlines the work on using IE libraries for extracting knowledge from the safeguarding reports. Section 3.3 presents initial approach for classification based on enriching feature vectors using WordNet. In Section 3.4, we investigate whether any existing lexical resources fit the safeguarding domain. Finally, Section 3.6 summarises findings and explains choices we made for following-up approaches.

3.1 Case Study: Safeguarding Reports

As already stated in Section 1.1 from Chapter 1, 'safeguarding' is a term used to denote 'measures to protect the health, well-being and human rights of individuals, which allow people, especially children, young adults, and vulnerable adults to live free from abuse, harm and neglect' [Quality Commission, 2014]. The safeguarding reviews are published by local authorities and community safety partnership. Their aim is to identify and describe related events that precede a serious safeguarding incident — for example, involving a child or vulnerable adult — and to reflect on agencies' roles and the application of current practices [Matters, 2006]. Each report contains key information about learning experiences and reflections on tackling serious incidents. The reports carry great potential to improve multi-agency work and help develop better safeguarding practices and strategies. Therefore, analyzing and understanding the safeguarding reports is crucial for health and social care agencies.

Depending on the type of crime committed there are four main types of safeguarding documents: Domestic Homicide Reviews (DHRs), Mental Health Homicide Reports (MHHRs), Adult Practice Reviews (APRs), and Child Practice Reviews (CPRs). However, we focus on DHRs which review the circumstances in which 'the death of a person aged 16 or over has resulted from violence, abuse or neglect from either a person to whom he or she were related or with whom he or she was or had been in an intimate personal relationship, or a member of the same household as him/herself' [Robinson et al., 2019].

3.1.1 Wales Safeguarding Repository

Despite the potential of safeguarding reports to help learn from previous tragic incidents, it is unclear the extent to which their findings have added to the sum of professional knowledge as previous analysis on these reports have been performed on a very small scale Robinson et al. [2019]. This can be attributed to the hard accessibility to

such reports and the lack of centralised point of access to the collection.

This is the motivation for the creation of WSR [1], conducted by CSRI, which involves a collaborative work between social scientists and computer scientists and aim to build a document repository for housing safeguarding reviews and reports. During the initial phase of the project, started in March 2018, a prototype repository, housing multiple types of safeguarding reports was created. Further, a coding framework was designed to help manual and automatic enrichment of the reports housed within the repository.

The main goals of the repository are:

- Improved accessibility — provide a single point of access to historical reports which will serve as a source of experience on good and bad practices on tackling and preventing serious crimes.
- Improved learning — develop techniques for extracting key lessons and main discussion topics from the reports which can serve as a guidance for preventing and tackling similar cases.
- Improved governance — involve identifying recommendations from reports which will help follow up on whether these recommendations are implemented and what is their effect on professional practice 'on the ground'.

The repository is in development stage and due to the sensitive nature of the documents, it is not available for external use. At its current status, the WSR provides only basic functionalities such as reports viewing and document search by name. An example of the interface of the repository is given in Figure 3.1.

Domain experts perform the thematic analysis manually. However, each report is lengthy and complex, so manual extraction of information is a time-consuming and potentially bias-prone process. Furthermore, in our particular case, the safeguarding

[1] WSR page: `http://upsi.org.uk/projects-2/wsr`

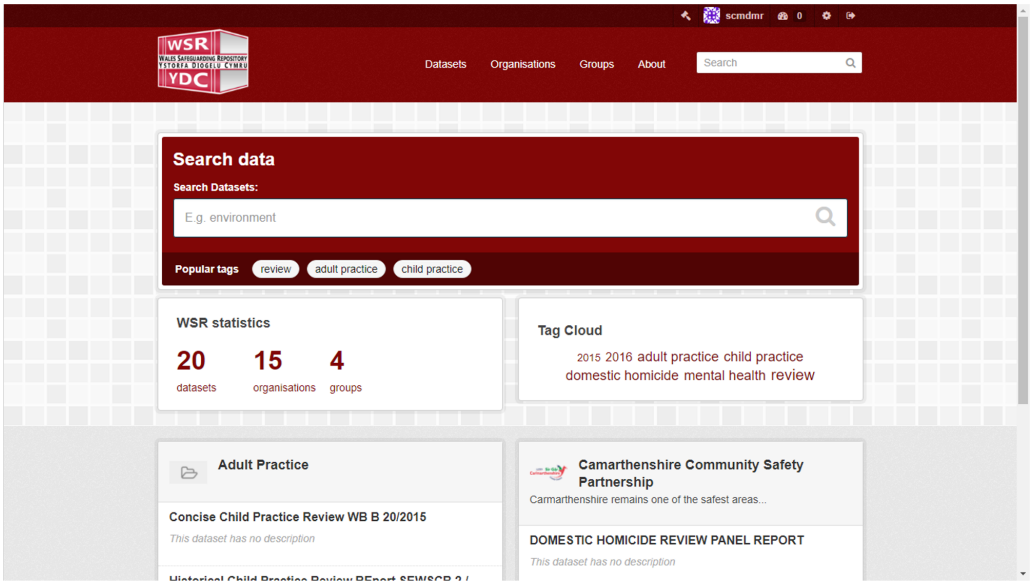

Figure 3.1: Wales Safeguarding Repository Interface.

collection is expected to grow significantly in the near future, with the additional resourcing of 500 historical reports, making the manual coding of these additional documents unfeasible. Therefore, the techniques developed throughout the thesis will be used to extend onto the functionality provided by the WSR and support automatic coding of incoming reports and thus provide more efficient searchability of the dataset. Developing these automated functionalities could help free up resources and assist practitioners from health and social care agencies in making faster and more accurate decisions (see Figure 3.2).

In the next stage of the project starting in May 2021, we will focus on incorporating classifier models developed during this thesis for detecting themes within the documents automatically and support search based on themes within the safeguarding collection.

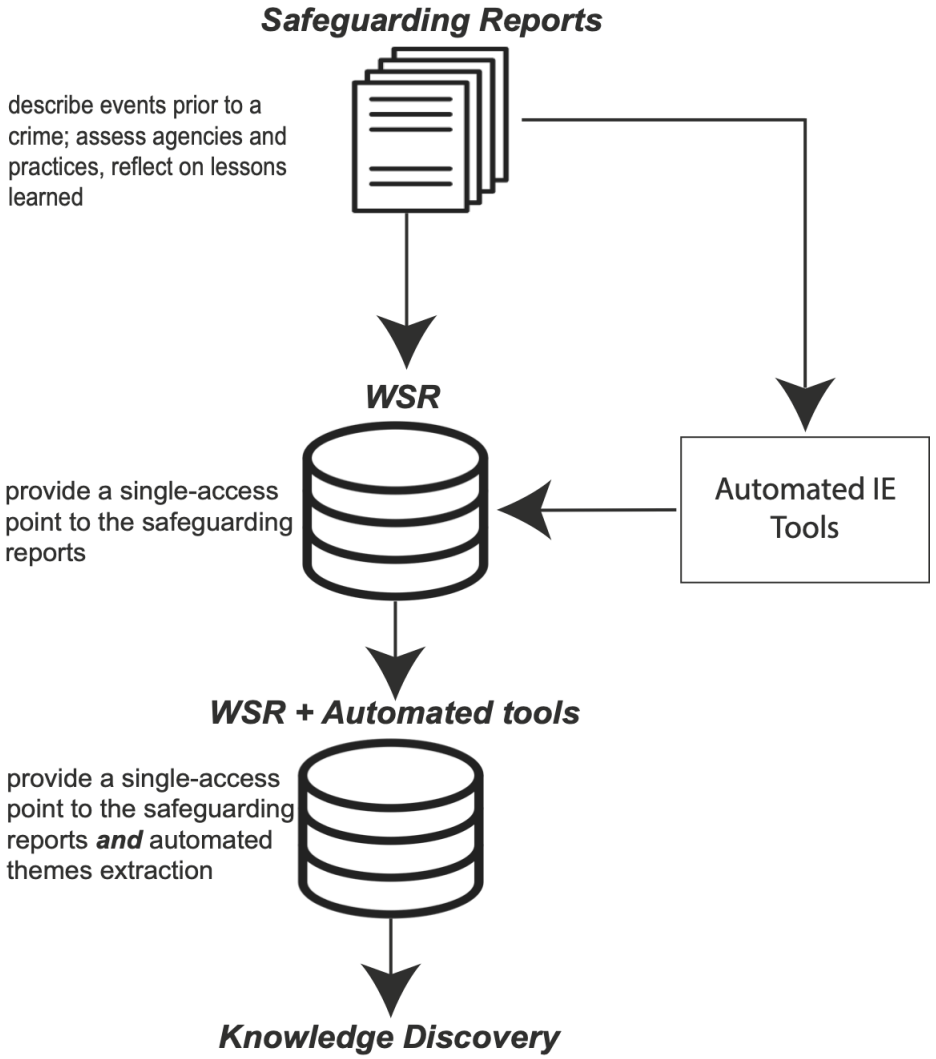

Figure 3.2: Wales Safeguarding Repository Workflow.

3.1.2 Thematic Framework

Traditionally, thematic analyses are done in social science by a process of manually 'coding' the reports: annotating them with themes identified by subject-matter experts. The process of coding documents and creating a thematic framework usually involves a multi-disciplinary research team where each member does the coding individually for each review, but the final version of the coding framework is outlined during regular discussions between team members Robinson et al. [2019].

Similarly, the thematic framework for the safeguarding reports resulted from collaborative work between multiple subject-matter experts. The initial thematic framework was heavily influenced by the findings of a thematic review looking across several safeguarding report types, presented by Robinson et al. [2019]. In this context, a *theme* refers to the main topic of discussion related to safeguarding incidents, specifically relevant to domestic homicide and mental health homicide. The documents were annotated with 5 overall themes (see Figure 3.3 and 96 sub-themes with different hierarchy depth (see Table 3.1) where a given part of the report can be coded with multiple themes. We refer to the coded sections of the reports as 'passages' throughout the thesis, where each passage can consist of a short phrase such as 'possession of drugs' or a list of sentences which could be viewed as short paragraphs.

```
                        levelName sentences passages
1  Safeguarding Reports Themes         NA       NA
2   ¦--Contact with Agencies         1616      485
3   ¦--Indicative Behaviour          1354      506
4   ¦--Indicative Circumstances       531      201
5   ¦--Mental Health Issues           392      134
6   ¦   ¦--Children                    10        4
7   ¦   ¦--Victim                     114       36
8   ¦   ¦--Perp                       136       43
9   ¦   °--Suicidal Ideation           90       30
10  °--Reflections                    983      341
```

Figure 3.3: Theme Hierarchy for the Safeguarding Reports.

There are 27 documents in total in the DHR collection, where each report has been coded within the five overall themes described in Table 3.1, which are 'Contact with Agencies', 'Indicative Behaviour', 'Indicative Circumstances', 'Mental Health Issues' and 'Reflections'. The themes 'Indicative Behaviour' and 'Indicative Circumstances' are similar in their semantics and purpose as they both describe problems that can serve as indication for the crime that was committed and can be used as signs by professionals to prevent similar crimes in future. However, 'Indicative Circumstances' is focused mainly on events which show relationship-related problems between the people involved in the crime and the events that have direct implications for committing the crime (i.e. the victim trying to get a divorce from the perpetrator, suspicion of infidelity). On the other hand, 'Indicative Behaviour' looks more in depth of personal

characteristics of the people involved in the crime that has been occurring throughout their lives, such as involvement in previous offences, signs of aggression, substance misuse and alcohol misuse. For giving more clarity over the differences between these two themes we show their sub-themes in Figures 3.4 and 3.5.

Theme	Description	Example
Contact with Agencies	Covers all agency interaction such as police contact, involvement with the third sector, contact with GPs and hospitals etc.	'The victim contacted Children's Services in September stating that she was struggling to cope with the children. This was progressed to an initial assessment undertaken by the Children With Disability Team.'
Indicative Behaviour	Describes the types of behaviour that might be indicative for the crime, such as signs of aggression, substance misuse, and previous offences, and disguised compliance	'His first conviction for assault was recorded when he was 15 years of age and he subsequently has had numerous episodes of detention/imprisonment along with a range of other penalties for the offences he has committed.'
Indicative Circumstances	'Describes the circumstances prior the incident such as relationship problems, debt, homelessness, and sex work	'The only mention of relationship problems occurred in 2010 when the victim mentioned relationship difficulties to her doctor. The perpetrator was present. The doctor referred them to RELATE but, so far can be ascertained, they never made an appointment.'
Mental Health Issues	Provides indications of any mental health problems that any of the involved people in the crime experienced	'A was then asked if any other services could have helped. A said that, with hindsight, he should have sought counselling and he was suffering from mild depression at the time. He had spoken with his family, but not medical services.'
Reflections	Discusses key lessons learned in reviewing the case. It covers failures/missed opportunities, family engagement, reports and re-organisation of public services	'Upon receipt of this information had Housing decided not to accommodate E, there would have been sufficient time prior to his release in June for alternative housing arrangements to have been considered. Linked to this is the wider issue about a lack of suitable accommodation for people who pose a high risk of harm to others. There are no specialist resources for such individuals available to housing. This review questions whether E's specific needs could have been met in more suitable provision and who has access to such provision, if it exists.'

Table 3.1: Overall themes description, where the examples given are non-verbatim examples of passages annotated with one of the five overall themes.

	levelName	sentences	passages
1	Safeguarding Reports Themes	NA	NA
2	°--Indicative Behaviour	1354	506
3	¦--Lying	57	17
4	¦--Weapons	583	170
5	¦--Emotional Abuse	7	3
6	¦--Self Inflicted Harm	17	5
7	¦--Stalking	30	11
8	¦--Offending	254	95
9	¦ ¦--Incidents resulting in NFA	38	9
10	¦ ¦--Theft and Kindred Offences	95	34
11	¦ ¦--Possession of Weapons	11	6
12	¦ ¦--Dangerous Driving or DUI	16	7
13	¦ ¦--Drug related crimes	17	8
14	¦ ¦--Manslaughter	5	1
15	¦ ¦--Offence against the Person	86	30
16	¦ ¦--Arson or threat thereof	32	11
17	¦ ¦--Fraud or Kindred Offences	2	2
18	¦ ¦--Assault on a Police Officer	11	4
19	¦ ¦--Offence Against the Property	41	14
20	¦ °--Public Order Offences	35	12
21	¦--Serious Threats to Life	64	22
22	¦ °--Previous attempts to kill	28	8
23	¦--Domestic Violence	372	104
24	¦ ¦--Victim	200	55
25	¦ °--Perp	182	50
26	¦--Substance Misuse	471	168
27	¦ ¦--Overdoses	41	21
28	¦ °--Drug Misuse	224	72
29	¦ ¦--Victim	13	7
30	¦ °--Perp	109	35
31	¦--Harrassment	57	20
32	¦ ¦--Victim	1	1
33	¦ °--Perp	52	17
34	¦--Controlling Behaviour	228	69
35	¦ ¦--Victim	23	8
36	¦ ¦--Perp	118	35
37	¦ °--Coercion	42	12
38	°--Aggression	364	98
39	¦--Children	7	2
40	¦--Victim	11	2
41	°--Perp	190	48

Figure 3.4: Indicative Behaviour Sub-themes.

	levelName	sentences	passages
1	Safeguarding Reports Themes	NA	NA
2	°--Indicative Circumstances	531	201
3	¦--Bereavement	32	10
4	¦ °--Victim	4	1
5	¦--NFA, Homelessness or Constantly changing Address	32	11
6	¦ ¦--Victim	17	6
7	¦ °--Perp	15	5
8	¦--Family Structure	94	34
9	¦ °--Seperation following Marriage	4	4
10	¦--Child Safeguarding	89	32
11	¦ ¦--Children engaging in Criminality	7	2
12	¦ °--Children exhibiting similar behaviours to parents	9	2
13	¦--Relationship Breakdown	36	15
14	¦--Debt or Financial Exploitation	74	27
15	¦ ¦--Victim	58	17
16	¦ ¦--Gambling	4	1
17	¦ °--Perp	7	5
18	¦--Sex Work	10	4
19	¦--Relationship with Children	82	28
20	¦--Relationship with Parent(s)	54	19
21	°--Quality of Relationship	87	34
22	°--Attempts to escape DV relationship	34	12

Figure 3.5: Indicative Circumstances Sub-themes.

The ‘Mental Health Issues’ theme is a separate overall theme from ‘Indicative Behaviour’ and ‘Indicative Circumstances’ in order to allow expansion of the framework for different types of safeguarding reviews such as the MHHRs.

In Table 3.2, we look at a distribution of both sentences and passages among the themes where sentences are obtained by splitting the coded passages into sentences and assigning each sentence the theme of the passage it belongs to. The data instances are unequally distributed among the themes, especially on a sentence level where the 'Contact With Agencies' is the best represented theme with 1616 sentences while the 'Mental Health Issues' theme is the worst represented with only 392 sentence followed by 'Indicative Circumstances' with 531. The corpus consists in total of 1261 passages which can be split into 3591 sentences.

Theme	**#passages**	**#sentences**	**#avg passage length**	**#avg sentence length**
Contact with Agencies	485	1,616	47	17
Indicative Behaviour	506	1,354	42	17
Indicative Circumstances	201	531	42	18
Mental Health Issues	134	392	48	19
Reflections	341	983	49	20
Total	1,261	3,591	45	18

Table 3.2: Overall themes statistics, where *'#passages'* refers to the total number of passages per theme, *'#sentences'* refers to the total number of sentences per theme, *'avg passage length'* refers to the average number of tokens per annotated passage, *'avg sentence length'* refers to the average number of tokens of per annotated sentence.

3.1.3 Lexical and Structural Characteristics of the Reports

The safeguarding reports are written in a free manner without a clearly pre-defined structure. They represent wide range of crimes, mental health, behavioural, and social issues. All these is an indication of a diversified language and structure across the reports, which makes the information extraction task very challenging especially given the small amount of data.

The reports are organised into sections and subsections which can vary across documents and types (an example of the structure of a report is given in Figure 3.6). How-

ever, there is a common pattern of content organisation of the reports which consists of three main parts:

- Generic information about the report type — It includes definitions and circumstances under which the given report is written. This part of the reports is excluded from the manual thematic analysis as it does not provide any knowledge on the specific case. A non-verbatim example paragraph of this part of the reports is: *'The Domestic Violence, Crimes and Victims Act 2004, establishes at Section 9(3), a statutory basis for a Domestic Homicide Review, which was implemented with due guidance on 13th April 2011. Under this section, a domestic homicide review means a review 'of the circumstances in which the death of a person aged 16 or over has, or appears to have, resulted from violence, abuse or neglect...'.*

- Descriptive part — It describes the events of the safeguarding case, often involving one or more crimes. An example paragraph of this part of the reports is: *'He was also known to be violent and his behaviour was very unpredictable. Domestic Violence against the Victim was not unusual and episodes of serious threats to the Victim's life were also known. Attempting to strangle her and attempting to drown her in a bath of water was the description of some of the attacks he made on the victim.'*

- Reflective part — It consists of findings: lessons learned and recommendations. An example paragraph of this part of the reports is: *'It is recorded that as far back as 1993, both the Perpetrator and the Victim were known to Mental Health Services. He presented with a number of problems including problematic alcohol use, depression, and aggressive behaviour. There were frequent references to domestic violence but opportunities to recognise the ongoing risk of violence were missed.'*

Figure 3.6: Report structure.

The reports are terminology-rich and contain highly-specialised language (see example of term frequency analysis on documents in Figure 3.7), which is also distinctive for the different themes. For instance, among the top terms for 'Contact with Agencies' are appointment-related terms and agencies such as 'psychiatric appointment', 'outpatient clinic', and 'gp practice' while 'Indicative Behaviour' is related to many offence-related terms such as 'common assault' and 'public order'.

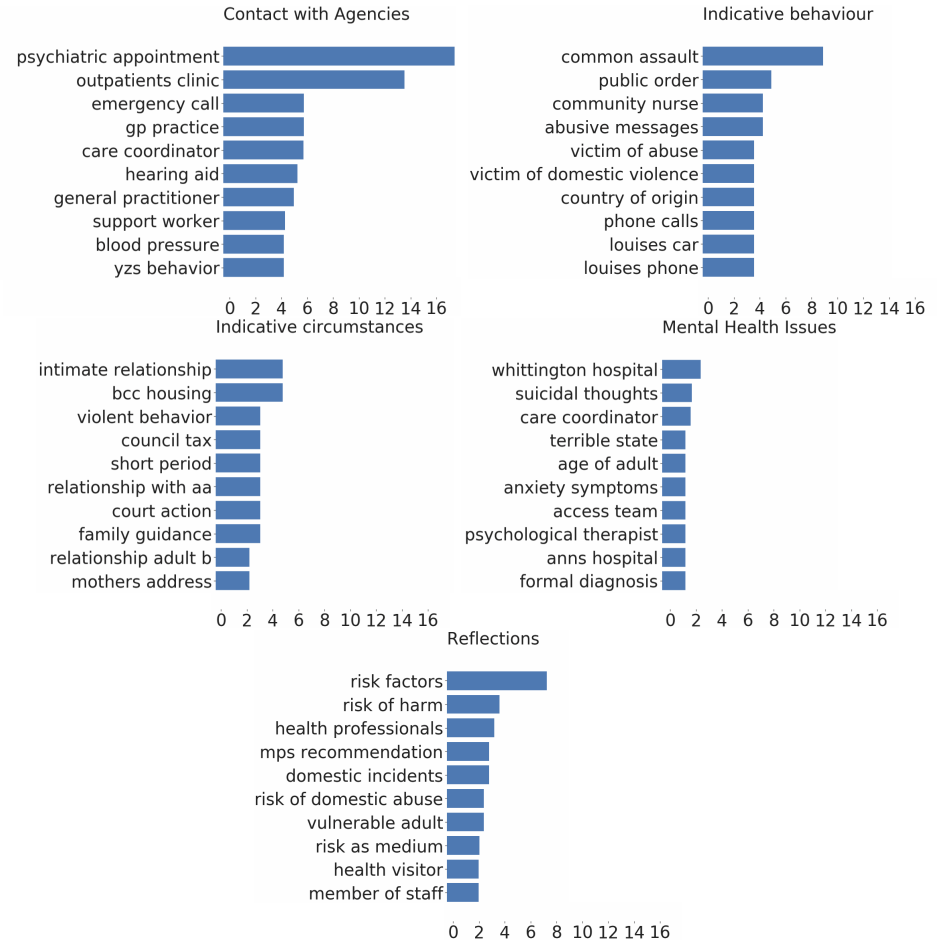

Figure 3.7: Terminology used in the reports based on TF-IDF, where multi-token terms are extracted from the corpus using FlexiTerm Spasić et al. [2013], an open-source software tool for automatic recognition of multi-word terms.

3.2 Information Extraction and Sentiment Analysis using Publicly Available Libraries

The analysis presented in this section are part of exploratory profiling of the dataset where we looked at what are the key features of the corpus in terms of actors such as agencies and also highlighting parts of the reports with a particular importance. For these purposes, we used well established tools for extracting named entities and for performing sentiment analysis. Specifically, we used entity extraction to identify key

features in terms of individuals, organisations and locations. Further, we used sentiment analysis for identifying sentences with key information. We hypothesised that key information would correspond to highly-emotive language and therefore, sentences with strong sentiment, e.g. positive or negative, might have a particular significance for the text.

3.2.1 Techniques Overview

We evaluated five off-the-shelf IE libraries which are Stanford Core NLP, Google Cloud API, GATE, SentiStrength, and NLTK. We have described these libraries in depth in Section 2.1.1 Chapter 2. One of the reasons for choosing these tools is that they use different approaches for performing NER and sentiment analysis which also cover the three main IE methods described in Section 2.1.1, Chapter 2. This allows identifying which tools are most suitable for performing IE tasks for the safeguarding domain. For instance, Stanford Core NLP and Google Cloud API are based on machine learning methods while GATE is using a hybrid approach of dictionary lookup and heuristics, and SentiStrength is based on dictionary lookup method (see Table 3.3). Additionally, these libraries have been used successfully for many IE tasks, including specialised domains [Chen et al., 2017, Maynard and Funk, 2020].

Tool	**Sentiment Analysis**	**NER**
Stanford Core NLP	Recursive Neural Tensor Network	CRF classifier
Google Cloud API	Deep learning models	Deep learning models
Gate	Generic Sentiment Analysis application	ANNIE dictionary look-up and rules
SentiStrength	dictionary look-up	NA
NLTK	NA	MaxEnt chunker

Table 3.3: Description of IE libraries used for performing NER and sentiment analysis.

3.2.2 Analysis

Evaluation Method

The performance of the tools for NER and sentiment analysis is compared against the annotations of three non-expert annotators. The study used a *description* and a *reflection* sets. Both sets consisted of 100 randomly-chosen sentences from the two main parts of the reports, i.e. 'descriptive part' and 'reflective part' as described in Section 3.1.3. The description set consisted of sentences describing the events of the safeguarding case — often involving one or more crimes — while the reflection set consisted of findings: lessons learned and recommendations. The two sets differed in the nature of how the sentiments of the sentences can be interpreted. The highlights of the descriptive set are the events; thus, the sentiment of the sentences will be judged by the sentiment of the event. An indicative (non-verbatim) example of a descriptive sentence is: 'Prison staff found the subject had hanged himself'. This sentence describes a negative event, i.e. a death. The highlights of the reflection sentences are the findings. Thus, the sentiment of the sentences will be judged by the sentiment of the comment. An indicative (non-verbatim) example of a reflective sentence is: 'The key finding from the review of the agencies' involvement is that there was strong evidence of good inter-agency working and appropriate referrals between local services'. This sentence express a positive reflection on inter-agency communication. We performed evaluation for the sentiment analysis using both, the *description* and the *reflection* sets because they differ in the way in which the sentiment can be interpreted. However, for the NER analysis we used only the *reflection* set.

Sentiment Analysis

Method: Sentiment scores, produced by text analysis tools have been normalised into three labels - 'positive', 'negative', and 'neutral'. This was done in order for the scores from the different text analysis tools to be comparable (see Table 3.4).

	positive	**negative**	**neutral**
Stanford Core NLP	positive, very positive	negative, very negative	neutral
Google Cloud API	score > 0.0	score < 0.0	score = 0.0
GATE	positive	negative	neutral
SentiStrength	score > 1	-5 $\leq$ score $\leq$ -1	score = 0

Table 3.4: Normalisation of IE libraries sentiment scores into *'positive'*, *'negative'*, and *'neutral'* labels.

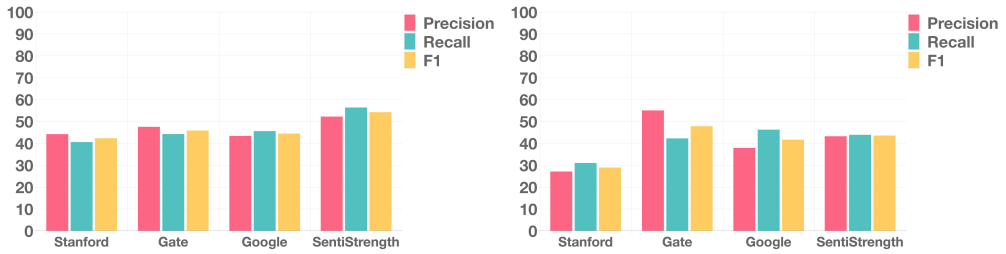

Figure 3.8: Average precision, recall, F1 for sentiment analysis: description set (left), reflection set (right).

Results: Figure 3.8 presents the average precision, recall, and F1 between the positive, negative, and neutral sentiment categories. These results show an unsatisfactory level of performance from the tools used. Overall, the tools performed better for descriptive sentences: SentiStrength performed the best for these with around 55% for precision, recall and F1. GATE performed best for reflective sentences with F1 of 48%. The poor performance of the tools can be attributed to the fact that they are trained on datasets very different to the safeguarding domain. For example, Stanford CoreNLP is trained on movie reviews where a phrase such as "with recommendation" has a positive sentiment while the same phrase in the context of a safeguarding report might have a negative sentiment (e.g. "sentenced to life imprisonment *with recommendation* of years"). SentiStrength is based on a dataset of MySpace content and uses a dictionary-based approach. It follows that sentences mentioning entities such as 'Specialist Dementia home' will match to the term dictionary 'special*' and thus have a positive sentiment.

Extracting Named Entities

Method: We extracted only entities belonging to the categories 'person', 'organisation', 'location' as people and organisations are the actors in these reports and locations help identify potential regions which can be associated with certain types of crime. Stanford Core NLP returns location-related tags such as 'city' and 'country'. These have been combined into the single tag 'location'.

Results: The results from the NER again show poor performance across all categories with F1 lower than 60%. Precision and recall tend to be very unbalanced. Some of the reasons for the high number of miss-classified entities are: anonymised individuals' names (e.g. 'victim', 'perpetrator', 'doctor 1'), and reviews and document names often classified as organisations (e.g. 'Adult Practice Review'). Specifically, F1 measure is surprisingly low for extracting locations with a tool such as Stanford. Examples of entities miss-classified as locations by Stanford CoreNLP are 'Wales Hospital', 'Huggard Centre', 'Greater', 'Wales'. Examples of locations that were missed by Stanford are 'Dyfed-powys', 'Bridgend', 'Greater Manchester Area'.

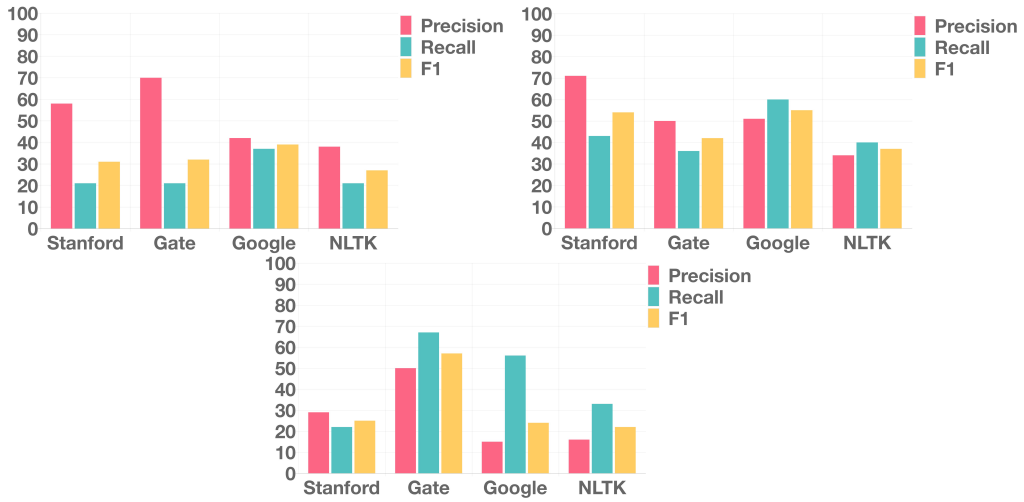

Figure 3.9: Evaluation results for entity extraction — person (left), organisation (middle), location (right).

Inter-human Agreement versus Inter-machine Agreement

We measured the inter-annotator agreement and the inter-tool agreement for our sentiment analysis and entity extraction exercises using Fleiss' Kappa Landis and Koch [1977]. The interpretation scale that we use for interpreting Fleiss' Kappa scores is given in Table 3.5.

K	Interpretation
<0	Poor agreement
0.0-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.0	Almost perfect agreement

Table 3.5: Interpretation of Fleiss' Kappa scores.

		Annotators	Tools
Sentiment analysis	Descriptions	0.6	0.1
	Recommendations	0.4	0.1
Entity extraction	Person	0.3	0.04
	Organisation	0.3	0.3
	Location	1.0	0.2

Table 3.6: Results from comparison between inter-human agreement and inter-machine agreement based on Fleiss' Kappa scores.

Fleiss' Kappa scores for the sentiment analysis showed good agreement between the annotators but a significant disagreement between the tools (see Table 3.6). The difference between the annotator scores for the two datasets suggests that humans find it easier to annotate the descriptive set rather than the reflective set while the tools did not differentiate between the two data sets. The vast majority of sentiment disagreement between the human annotators involved distinguishing between neutral versus positive/negative polarity. There was only a single instance of disagreement between positive versus negative polarity of a sentence: 'The person disclosed at an appointment, that they had overdosed a month before and now felt stupid about it' (this example is

paraphrased). However, the disagreement between the tools in terms of positive versus negative sentiment was considerably higher: 34% for the description and 36% for the reflections. The Fleiss' Kappa scores are low across all entity extraction categories for the software tools. Inter-human agreement for person and organisation are also low. Entities that humans disagreed on were: 'GP' (general practitioner), 'Coroner', 'Mental Health Teams', and 'Mental Health Tribunal', all of which tended to be labelled either as 'person' or 'organisation'. The entities on which human annotators disagreed can be considered as both organisation or person depending on the context and the purpose of the text. This shows that the annotation of specialised documents may require a prior training or agreement on principles of how entities are classified. However, these analysis helped identify that the entity extraction task is challenging not only for software but also for non-specialist human annotators. Further, the results from this study helped identify challenges in performing information extraction for the safeguarding documents.

3.2.3 Summary

The results from the sentiment analysis provide no evidence that off-the-shelf sentiment analysis tools can identify key parts of the safeguarding reports. Further, the unsatisfactory results of the entity extraction tools show the need for potentially more domain-targeted approaches.

3.3 Classification Augmentation with WordNet

In this section, we further extend on the exploratory work on how traditional IE approaches perform for the safeguarding domain by focusing on classification and feature augmentation strategies. Specifically, we augment the features extracted from the training set of the safeguarding reports using WordNet [Miller, 1998] before turning them

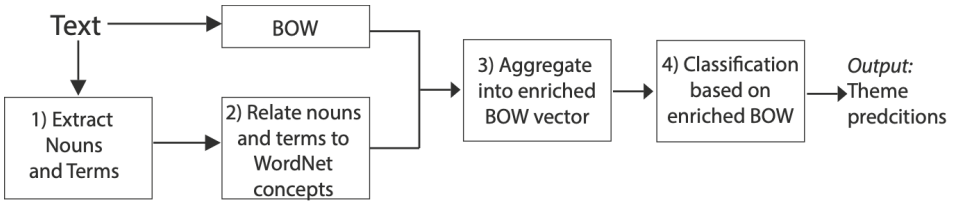

Figure 3.10: WordNet-based feature augmentation approach.

into vectors used by machine learning algorithms. We evaluate these augmentations using linear classification models. WordNet is a large human-constructed semantic lexicon of English words structured as a graph. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between synsets such as hypernyms, hyponyms, meronyms. For example, the word 'dog' is a *synonym of canine*, a *hypernym of puppy* and a *hyponym of animal*. The reason for choosing WordNet is the wide coverage of English words, the graph-like structure of the resource and its successful previous use to enhance word vectors Faruqui et al. [2015], Mrkšić et al. [2017] and classification performance. A popular approach for enriching feature vectors is through the use of word embeddings. However, we want to establish whether more traditional classification enhancement approaches involving lexical resources are suitable for the safeguarding domain as they are considered less resource consuming (explained further in Section 2.3.7). The methodology is similar to the one presented in [Heap et al., 2017]. However, we use WordNet for enriching feature representations rather than word models. The WordNet-based feature augmentation approach consists of four steps 3.10.

1) Extract nouns and terms: We used FlexiTerm [Spasić et al., 2013], as in Section 3.1.3 for extracting multi-token words. Further, we performed POS tagging using Stanford CoreNLP in order to identify nouns in the training corpus.

2) Relate nouns and terms to WordNet concepts: We adapted Lesk algorithm [Banerjee and Pedersen, 2002] for preforming word sense disambiguation (WSD) using Word-

Net as our sense inventory. The implementation of the algorithm is provided by [Tan, 2014]. Adapted Lesk [Banerjee and Pedersen, 2002] uses WordNet to infer the most suitable synset for a given word based on their neighbours. It creates a window of context by including a target word and the surrounding words. It compares glossaries between each pair of words in the context of window. Then, it selects the combination of glossaries with the highest overlap and assigns the target word the sense given in that combination. Our approach for associating related tokens is as follows: For each synset returned, we obtain the associated hypernyms and synonyms. Then, we associate nouns and terms with each other based on the following principle: if a token is contained within the hypernyms or synonyms of another token, then they are related. We assign a main token to each group of associated tokens based on their occurrence frequency. Initially, we performed experiments with relating tokens based on various combinations of synonyms, hypernyms, and hyponyms. However, an observation of the different relations produced with each one of the methods showed that hypernym and synonym-based relations introduce less noise.

3) Aggregate into enriched BOW vector: We build sentence vectors by replacing the tokens that are associated with the main token of each association group.

4) Classification based on enriched BOW: We use the enriched vectors for performing classification. We use two widely used classification models, GNB and SVM. Despite their simplicity, these models provide strong performance for many text classification tasks [Wang and Manning, 2012, Islam et al., 2007], and SVM has been successfully used for evaluating feature enrichment methods for smaller and more specialised domains [Gazzotti et al., 2019, 2020, Colace et al., 2014]. Further, GNB provides a native description of the decision which has been used in an initial stage for adjusting a stop words list after discussion of the interpretations with a domain-expert.

Further, it is possible to provide a native interpretation of the decision of these algorithms. We used the algorithms implementation available in the scikit-Learn lib-

rary [Pedregosa et al., 2011] with default settings.

3.3.1 Dataset

We performed analysis on a sentence level where each sentence is given the label of the passage to which it belongs to. We evaluated classifiers using a development and a test set. The train and development sets were extracted from the original coded documents, discussed above, while the test set was part of safeguarding reviews which were not previously coded by experts. The test set contained 100 randomly selected passages where each passage consisted of 3 sentences. These passages were annotated by one of the experts who participated in the creation of the thematic framework. The test passages which were not given a label by the subject-matter expert as they were considered not informative enough for a label, are excluded from the test set. The dataset statistics are given in Table 3.7

Theme	**#train**	**#dev**	**#test**
Contact with Agencies	1,281	335	219
Indicative Behaviour	1,078	276	83
Indicative Circumstances	427	104	99
Mental Health Issues	316	76	51
Reflection	780	203	78
Total	**2,736**	**685**	**284**

Table 3.7: Overview of the safeguarding dataset, where *'#train'* refers to the number of training sentences, *'#dev'* refers to the number of sentences used in the development set and *'#test'* refers to the number of sentences used as test set.

3.3.2 Evaluation

We evaluated the performance of the machine learning algorithms using precision, recall, and F1-measure metrics. The average results are calculated by using micro-precision and micro-recall (3.1), and macro-precision and macro-recall (3.2). The

micro- measures are based on the average of the number of true positives, false positives, and false negatives while the macro- measures are based on averaging precision and recall between the classifiers performance for each of the themes. Therefore, the micro-average gives a better understanding of how the system performs overall while the macro-average is more sensitive to class imbalances. We use these evaluation measures throughout the thesis.

$$micro - p = \frac{\sum_{i=1}^{n} tp}{\sum_{i=1}^{n} tp + \sum_{i=1}^{n} fp} micro - r = \frac{\sum_{i=1}^{n} tp}{\sum_{i=1}^{n} tp + \sum_{i=1}^{n} fn} \tag{3.1}$$

$$macro - p = \frac{\sum_{i=1}^{n} precision}{n} macro - r = \frac{\sum_{i=1}^{n} recall}{n} \tag{3.2}$$

3.3.3 Classification Results

Results from classification are presented in Table 3.8.

SVM classifier performance is lower than the performance of GNB classifier and thus we have presented only SVM results for 1,2 grams (see Table 3.8). WordNet-based feature enrichment leads to higher recall for almost all themes except 'Mental Health Issues'. However, the precision of 'GNB-WordNet' is much lower than the precision values for the baseline method. Therefore, micro- and macro- average results are always lower for 'GNB-WordNet' approach in comparison to the baseline (i.e. GNB - 1,2 grams). Problems with using a generic lexical resource for the safeguarding reports include the lack of relations between semantically similar words and the existence of irrelevant relations within the context of the safeguarding reports (see Table 3.9). Examples of terms and tokens that are not related by WordNet are *'domestic violence'*

		Dev set			Test set		
Method	**Theme**	**p**	**r**	**F1**	**p**	**r**	**F1**
GNB - 1,2 grams	Contact with Agencies	.650	.700	.680	.860	.470	.610
	Indicative behaviour	.560	.630	.590	.460	.570	.510
	Indicative circumstances	.330	.640	.440	.520	.510	.510
	Mental Health Issues	.260	.570	.360	.390	.450	.420
	Reflections	.580	.690	.630	.470	.760	.580
Macro-Average		**.480**	**.650**	**.540**	**.540**	**.550**	**.550**
Micro-Average		**.510**	**.660**	**.570**	**.550**	**.570**	**.530**
SVM - 1,2 grams	Contact with Agencies	.680	.620	.650	.770	.360	.490
	Indicative Behaviour	.720	.360	.480	.680	.280	.390
	Indicative Circumstances	.990	.080	.160	.990	.020	.040
	Mental Health Issues	.670	.020	.050	.350	.020	.001
	Reflections	.810	.310	.450	.700	.240	.360
Macro-Average		**.770**	**.280**	**.360**	**.630**	**.180**	**.260**
Micro-Average		**.710**	**.370**	**.490**	**.740**	**.230**	**.350**
GNB - wordNet	Contact with Agencies	.520	.860	.650	.750	.790	.770
	Indicative Behaviour	.470	.790	.590	.360	.690	.470
	Indicative Circumstances	.280	.620	.390	.330	.320	.330
	Mental Health Issues	.190	.460	.270	.200	.240	.220
	Reflections	.370	.750	.500	.280	.620	.390
Macro-Average		**.370**	**.700**	**.480**	**.390**	**.530**	**.440**
Micro-Average		**.410**	**.760**	**.530**	**.450**	**.600**	**.510**

Table 3.8: Classification results using statistical-based feature vectors and WordNet-based augmentation, where *'Dev set'* refers to development set, *'p'* refers to precision, and *'r'* refers to recall, and *GNB* refers to Gaussian Naive Bayes classifier.

and 'physical abuse', *'substance misuse' and 'drug abuse'*, *'alcohol misuse', 'drinking problem', and 'alcohol abuse'*. Another problem is the irrelevant relations produced by WordNet between some tokens. Examples include *'misuse' related to 'use' and 'exercise'* and *'substance' related to 'content', 'message', 'guidance', 'counselling'*. Some examples of good token relations extracted from WordNet is for the word *'schizophrenia' - 'schizophrenic psychosis', 'dementia praecox', 'schizophrenic disorder', 'psychosis', 'paraphrenia', 'paranoid schizophrenia', 'psychosis'*.

problem	example
lack of relations	'domestic violence' and 'physical abuse'; 'substance misuse' and 'drug abuse'; 'alcohol misuse', 'drinking problem', and 'alcohol abuse'
irrelevant relations	'misuse' related to 'use' and 'exercise'; 'substance' related to 'content', 'message', 'guidance', 'counselling'

Table 3.9: Examples of problems with WordNet relations.

Overall, results from this experiments showed that generic publicly available lexical resources do not address the needs of the safeguarding domain.

3.4 Investigations into Lexical Resources

The findings from Section 3.3 show that generic lexical resources are unsuitable for the specialised dataset. Therefore, we extend investigation into existing lexical resources and datasets that are more related to the safeguarding domain. We searched through three purpose-build search engines, Swoogle [2] (last accessed on 01/03/2019) and The Linked Open Data Cloud [3] (last accessed on: 13/12/2020) for ontology search, and Google dataset search [4](last accessed on: 13/12/2020). We used 9 search terms descriptive of the safeguarding reports (see Table 3.10). We perform search for datasets related to the safeguarding domain in addition to suitable knowledge graphs, because additional unlabelled data can still help enhance classification for the safeguarding reports.

The active ontologies found during this initial research were considered unsuitable due to their unfinished or generic structure, i.e. containing only a few terms. Further, the datasets found from Google Dataset Search represent mainly structured, categorical, and statistical data which are not complementing the safeguarding reports. Fur-

[2] http://swoogle.umbc.edu/2006/

[3] https://lod-cloud.net

[4] https://datasetsearch.research.google.com

Search term	Swoogle		The LOD Cloud		Google Dataset Search		
	#returned results	#active kg	#returned results	#active kg	datasets	#returned results	#active kg
'crime'	99	11	5	1	100+	0	0
'safeguarding'/'safe guarding'	7	0	0	0	100+	0	0
'vulnerable adult(s)'	0	0	0	0	100+	0	0
'homicide'	0	0	0	0	100+	0	0
'murder'	40	17	0	0	100+	0	0
'healthcare'	0	0	2	2	100+	0	0
'mental health issues'	0	0	0	0	100+	0	0
'mental health'	5	2	0	0	100+	0	0
'psychiatric'	20	4	0	0	100+	0	0

Table 3.10: Results returned from search engines for availability of knowledge graphs related to the safeguarding domain, where *'#returned results'* refers to total number of returned results per search term and *'#active kg'* refers to number of active knowledge graphs returned per search term.

ther, none of these datasets were associated with knowledge resources. This implies the need for creating a safeguarding knowledge graph. However, creating a domain-specific knowledge graph requires a continuous input from domain experts which are sparse for the safeguarding domain and it can potentially limit the usability of methods to domains different from the safeguarding. Therefore, creating a safeguarding knowledge graph is outside the scope of the thesis.

3.5 Discussion

3.5.1 Information Extraction using Publicly Available Libraries

A wide range of IE libraries were evaluated in order to explore different approaches for NER and sentiment analysis. As these were only exploratory analysis into applicability of existing methods, we did not adapt them to the domain. The low performance of tools and the high level of disagreement according to Kappa score between tools and human annotators showed that analysing the safeguarding reports is a challenging task even for non-expert humans.

3.5.2 Suitability of Lexical Resources for the Safeguarding Domain

The use of publicly available lexical resources for enriching feature vectors is a widely used method, especially for more specialised domains such as medical domain, for improving classification performance. The simplicity of this model, as it does not require the creation of knowledge graph, and the low resource requirements of statistical machine learning models are the main advantages of such an approach. However, the evaluation showed that a generic lexicon such as WordNet is unsuitable for the safeguarding domain. A more targeted research into existing knowledge graphs and datasets that can supplement the corpus showed a lack of more specialised resources fitted to our needs.

3.6 Conclusions

In this chapter, we described the case study of the safeguarding reports which we use throughout the thesis for testing our hypothesis. We outlined our practical goal which is automating the thematic analysis for the safeguarding domain. Previously, thematic analysis have been performed manually. However, with the growing collection of the reports, manual annotation becomes cost- and time- consuming, and therefore unfeasible approach. The automation of the thematic analysis will free up resources and provide more efficient searchability through the collection for practitioners and researchers. Additionally, the small volume of documents, the highly specialised and diverse lexicon and structure make the application of IE approaches, especially supervised machine learning, a challenging task. The lack of extensive research into developing, and extensively researching supervised classification methods for such specialised documents with limited dataset, is the main motivation for this thesis.

In this chapter, we also presented our initial analysis on the applicability of more traditional IE approaches for NER and a simple approach for augmenting classification

performance using publicly available resources. Findings from this exploratory work showed the need for using more domain-targeted and contextually-aware approaches.

Question **RQ1** from Section 1.2 has been addressed in order to investigate whether generic lexical resources can facilitate information extraction from specialised texts such as the safeguarding reports. Results from this exploratory work showed that extracting knowledge from the safeguarding domain potentially requires more domain-targeted or contextually-aware approaches. The investigation into suitable lexical resources which can be used to facilitate the prediction of themes showed lack of suitable resources for our domain. These lead to the need of either creating lexical resources that fit our needs or exploring more context aware approaches such as state-of-the-art neural models for performing classification. This motivated our following up research into more contextualised neural models which can be targeted to the domain without the need to create domain-specific lexicons. As the latter approach requires significant input from domain experts and it potentially limits the usability of methods to the safeguarding domain.

Chapter 4

Evaluation of State-of-the-art Classification Methods for the Safeguarding Domain

In the previous chapter, we evaluated a more traditional classification approach based on the use of linear classifiers and lexical resources widely applied for more domain-specific corpora. Results, however, proved such a naive but less resource consuming approach unsuitable for specialised texts such as the safeguarding reports and showed the need of using more context-aware methods. In this chapter, we extend the analysis on existing classification approaches but with a focus on state-of-the-art methods. Specifically, we explore various neural model-based approaches for feature extraction, feature integration, and classification. As a baseline for comparing the different approaches, we use a simple linear classifier coupled with count-based features.

In Section 2.3.6 we presented recent research on performing low resource text classification. The approaches were mainly based on using traditional neural models such as CNN and LSTM-based classification as well as exploring techniques such as fine-tuning for adapting more recent transformer-based models to the task. However, such approaches are limited to evaluation on generic datasets, similar to those used for pre-training neural network models and assume access to large amounts of unlabelled data. Further, there is no extensive evaluation on what combination of embedding and language models and classification approaches fit the needs of a small domain-specific

and terminology-rich corpus.

In this chapter, we address these gaps by comparing the performance of three supervised approaches - simple linear classification models, and text classification methods based on pre-trained and corpus-trained word embeddings, and state-of-the-art language models. The reminder of this chapter addresses question **RQ 2: Which classification approaches help preserve subject-matter expert knowledge for annotating specialised unstructured texts, compared to human annotators?** from the research questions identified in Section 1.2. More specifically, contributions include a thorough analysis of different methods and models for obtaining word and sentence representations and how these affect the classification task. In particular we look at how domain-trained embeddings compare to larger state-of-the-art language models and how deep learning approaches are affected by training dataset size versus the amount of context given.

Further, considering the subjective nature of the thematic analysis, we compare the machine learning algorithms performance to the annotations of domain experts who have not participated in the creation of the thematic framework. We refer to these domain experts as 'expert validators' and we use their annotations as a way of measuring how well classification approaches preserve the knowledge provided by the original annotator.

4.1 Eliciting Subject-matter Experts Opinion

As early experiments based on traditional approaches presented in Section 3.3 showed an imbalance between precision and recall of the classification models, we elicited expert preferences on the importance of the two measures. A survey was distributed among a closed group of subject-matter experts, six in total, all drawn from the target user group: analysts of safeguarding reports from our team of social scientist collaborators. The survey consisted of six questions, five of which were multiple-choice

with the sixth being free text for justifying responses. In the beginning of the survey we gave an explanation and example of precision and recall measures. Each of the multiple-choice questions represented two hypothetical options of system outputs where the outputs consisted of a number of relevant and irrelevant retrievals (see Figure 4.1). The numbers of relevant vs irrelevant retrievals were distributed among the two options so one option always had a higher precision and the other option had higher recall. A summary of the multiple-choice questions with corresponding precision and recall values are given in Table 4.1.

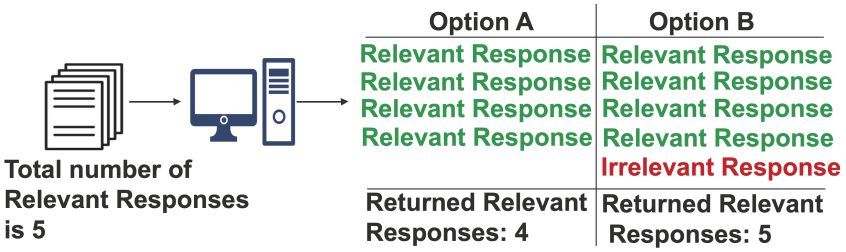

Figure 4.1: Example of the first question in the survey.

	Option A			Option B		
	retrievals	**p**	**r**	**retrievals**	**p**	**r**
Q 1	5 relevant & 2 irrelevant	0.7	1.0	3 relevant & 0 irrelevant	1.0	0.6
Q 2	4 relevant & 2 irrelevant	0.67	0.8	3 relevant & 1 irrelevant	0.75	0.6
Q 3	2 relevant & 2 irrelevant	0.5	0.4	4 relevant & 4 irrelevant	0.5	0.8
Q 4	4 relevant & 0 irrelevant	1.0	0.8	5 relevant & 1 irrelevant	0.8	1.0
Q 5	1 relevant & 0 irrelevant	1.0	0.2	4 relevant & 4 irrelevant	0.5	0.8

Table 4.1: Survey questions where *'Q'* stands for Question.

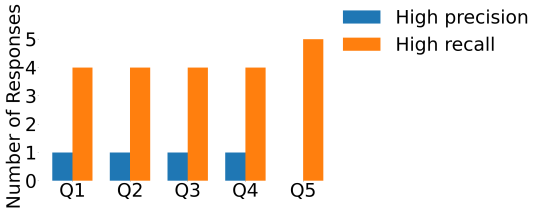

Figure 4.2: User survey results.

The results (see Figure 4.2) show that 83% of the respondents selected the system output associated with higher recall for each one of the questions. Respondents justified their choices by indicating that they preferred for the system to return more of the relevant responses even if this meant that more irrelevant ones would be returned as well. Even for questions (see Table 4.1, Q 4) where the number of relevant and irrelevant responses was very similar between the two options, respondents preferred completeness of responses rather than higher precision. Only one of the six respondents preferred higher precision over recall, and only then except in cases when the recall was very low (i.e., Option A for Q 5). This means that all participants were satisfied with precision above 50% and high recall, while a recall below 50% was deemed unacceptable even when the precision is very high.

4.2 Classification Methodology

The classification methodology we follow consists of three main steps, representing the main steps of the text classification process, outlined in Section 2.3. First, we obtain word vectors using pre-trained word embedding vocabularies, models built using the safeguarding collection, and pre-trained contextualised models. Then, we represent sentences using two approaches. The first approach is based on performing simple combination of the word embeddings while the second approach uses built-in sentence encoders. Finally, for performing classification we experiment with simple linear classifier, shallow neural network, and fine-tuned language model.

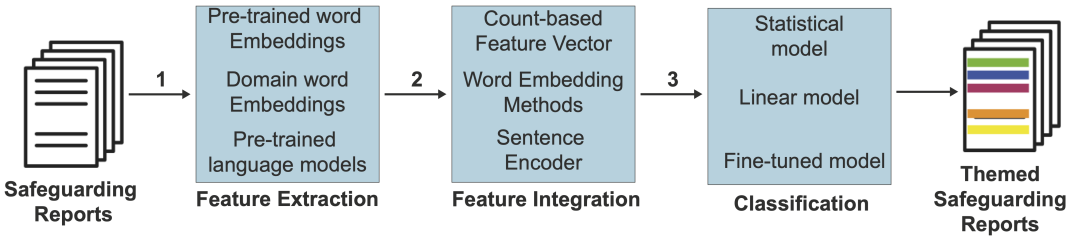

Figure 4.3: Methodology overview.

4.2.1 Feature Extraction

Pre-trained Word Embeddings The investigation from Section 2.1.4 in the Background chapter showed that the two most efficient and well established methods for obtaining word vectors are CBOW and skip-gram. Further, Word2Vec and fastText are the word embedding models based on these methods and have been successfully applied in many domains. Therefore, we leverage Word2Vec model, trained on Google news dataset and fastText trained with sub-word information on Common Crawl. A limitation of Word2Vec is that it ignores the morphology of words by assigning a distinct vector to each word. This limitation is addressed by fastText, where each word is represented as a bag of character n-grams. This allows fastText to build vectors for rare words, misspelled words or concatenation of words.

Corpus-specific Word Embeddings In order to learn domain-specific word embedding model we used the safeguarding reports corpus without including the part of the documents we use for evaluating classifiers. We use the same split between train, dev, and test data as we did in Section 3.3 for evaluating WordNet enhanced classification method. We use fastText for learning the embeddings because it captures the meaning of rare words better than other approaches. We use the skip-gram method for building word embeddings with 300 dimensions. Further, following the approach of [Yin and Schütze, 2016], we build metaembeddings by averaging corpus-trained and pre-trained word embeddings for each word in the safeguarding vocabulary in an attempt to improve sentence representations.

Pre-trained Language Model As mentioned in Section 2.1.4 in the Background chapter, a limitation of the word embedding models described above is that they produce a single vector of a word despite the context in which it appears. Therefore, we also include in our analysis BERT model, trained on Books Corpus and English Wikipedia, which generates more contextually-aware word representations.

4.2.2 Feature Integration

We use several ways for combining the word embeddings into reduced sentence representations, similar to [Li et al., 2018], where authors evaluate the effect of a range of word embedding and sentence embedding methods over the classification of tweets. Their broad experiments with various sentence encoder methods is one of the reasons for using their classification methodology as an example. Similar to them, we want to evaluate the performance of various well established embedding models and find which are the most suitable for a specialised domain, such as the safeguarding. Further, we also perform classification for short texts, i.e., sentences. However, we perform more extensive analysis by considering multiple types of classifiers and look at models performance for different data sizes versus the context given.

The first feature integration approach we consider is based on reducing dimensionality by applying arithmetic methods over the word embeddings while the second approach is based on using sentence encoders. We also use a simple count-based feature vectors for baseline approach.

Word Embedding Methods Despite the simplicity of word embedding-based methods, applying arithmetic functions over the word embedding vectors have proven to increase performance, especially when creating metaembeddings (see Section 2.2). This motivated us to experiment with the following three simple strategies for building sentence representations from word embedding:

- Mean — We average the embeddings of each word in a sentence along each dimension. Thus, a sentence vector will have the same dimension as a word vector/embedding.
- MinMaxMean (MMM) — In addition to mean, we also take the minimum and maximum over all the words in a sentence, along each dimension of the word vectors. Each aggregation, min/max/mean, will produce a vector that has the

same dimension as the word vectors. We concatenate the vectors corresponding to min/max/average, respectively, and obtain a tweet vector whose dimension is three times the dimension of the word vector.

- TF-IDF-Mean — We assign TF-IDF weights to the words in a sentence, and calculate the weighted average of the word embeddings along each dimension (where the contribution of a word is proportional to its TF-IDF weight).

Sentence Encoders We perform experiments with recent approaches to sentence-level encoding. We chose sentence encoders which follow different methodologies for building sentence representations and have also been widely applied in previous research.

We use unsupervised Smooth Inverse Frequency (uSIF) [Ethayarajh, 2018]. This method takes the weighted average of the word embeddings modified with Singular Value Decomposition (SVD) for dimension reduction. Further, it does not require parameter tuning. We also use InferSent [Conneau et al., 2017], a sentence embeddings method that is trained on natural language inference data. We chose this method because it generalizes well to many different tasks. The architecture is based on BiLSTM with mean/max pooling. For our experiments we use max pooling. Further, this model is build using earlier neural network models (see Section 2.1.3). We also use BERT [Devlin et al., 2019] as this model gives state-of-the-performance for many tasks. There are two steps in the BERT framework — pre-training and fine-tuning. In this step of the methodology, we use the base pre-trained BERT model, trained on the Books Corpus and English Wikipedia, for extracting contextualized sentence embeddings [Devlin et al., 2019]. The fine-tuning step consists of further training on the downstream tasks. By including InferSent and BERT in our analysis, we perform a comparison between a NN model based on LSTM and a transformer-based model.

4.2.3 Classification

We perform classification on a sentence level where each sentence had been assigned the theme of the passage the sentence belonged to. Here, we take 'ground truth' to be the codes produced by the expert annotators who were involved in the creation of the thematic coding framework (see Section 3.1.2). As mentioned in the beginning of this chapter, we use three types of classifiers where each classifier represents one of the main text classification methods outlined in Section 2.3.4 as part of the Background research. In this way we want to ensure a coverage of main existing state-of-the-art approaches.

Count-based Model We used the GNB algorithm available in the scikit-Learn library [Pedregosa et al., 2011] as a representative of a traditional linear classifier. We opted for this machine learning algorithm, since it performs better than SVM classifier as shown in previous chapter. Also, using the same simple classifier allows comparison between traditional approaches, presented in previous chapter, and state-of-the-art methods.

Linear Model Investigations in Section 2.3.4 outlined a potential problem with linear classifiers, i.e., they struggle with out-of-vocabulary (OOV) words, fine-grained distinctions and unbalanced datasets. The fastText classifier addresses this problem by integrating a linear model with a rank constraint, allowing sharing parameters among features and classes. This enables good prediction accuracy in classification tasks where some classes have very few examples. The model learns embeddings for each word in a sentence. This word representations are then averaged into a text representation, which is in turn fed to a linear classifier. We used default parameters and 'ova' as the loss function. Further, we defined a threshold of 0.4 for assigning a label to a given instance since we perform multi-label classification.

Fine-tuned Model We fine-tune BERT for the classification task using a sequence classifier, a learning rate of 5e-5 and 4 epochs. We use sequence classification, because of the sequence-like patterns that appear in the dataset. In particular, we adapted BERT to multi-label classification using the BERT's Hugging Face default transformers implementation for classifying sentences [Wolf et al., 2019] and [Rajapakse, 2019]. In order to tune the classifiers to produce higher recall we define a threshold of 0.4 for assigning a label to a given instance.

4.3 Experimental Results

4.3.1 Dataset Summary

We use the same dataset and train, dev, and test data distribution as in the previous chapter, Table 3.7 from Section 3.3.1. However, for clarity reasons, we also include dataset statistics in Table 4.2.

Theme	**#passages**	**#sentences**	**#avg passage length**	**#avg sentence length**
Contact with Agencies	485	1,616	47	17
Indicative Behaviour	506	1,354	42	17
Indicative Circumstances	201	531	42	18
Mental Health Issues	134	392	48	19
Reflections	341	983	49	20
Total	1,261	3,591	45	18

Table 4.2: Overall themes statistics, where *'#passages'* refers to the total number of passages per theme, *'#sentences'* refers to the total number of sentences per theme, *'avg passage length'* refers to the average number of tokens for coded passage, *'avg sentence length'* refers to the average number of tokens of a coded sentence (the same as Table 3.7).

4.3.2 Evaluation Metrics

We evaluate the performance of the machine learning algorithms by using the same metrics as in the previous chapter, i.e., precision, recall, and F1- measure metrics. The summary results are calculated using micro-precision and micro-recall, and macro-precision and macro-recall [Yang, 1999].

4.3.3 Overall Results

Method			**Micro-measures**			**Macro-measures**		
Classifier	**FE**	**FI**	**p**	**r**	**F1**	**p**	**r**	**F1**
Baseline	1,2 grams	count	.510	.660	.570	.480	.650	.540
GNB	Word2Vec	mean	.330	.530	.410	.360	.580	.400
		TF-IDF	.260	.480	.330	.320	.580	.330
		MMM	.370	.580	.450	.370	.600	.430
		uSIF	.320	.450	.370	.340	.470	.350
	fastText	mean	.380	.540	.440	.380	.550	.420
		TF-IDF	.270	.480	.340	.320	.560	.340
		MMM	.380	.560	.450	.380	.580	.430
		uSIF	.350	.540	.420	.360	.540	.400
		inferSent	.280	.650	.400	.290	.670	.390
	average	mean	.390	.520	.450	.390	.530	.430
		uSIF	.450	.590	.510	.440	.590	.480
	domain	mean	.470	.600	.530	.450	.610	.500
		uSIF	.440	.570	.500	.430	.590	.480
	Base BERT	BERT	.430	.600	.500	.400	.590	.470
fastText	domain	mean	.520	.670	.590	.480	.620	.540
	generic	mean	.520	.640	.570	.480	.590	.520
fine-tuned BERT	BERT	BERT	**.560**	**.730**	**.640**	**.520**	**.680**	**.590**

Table 4.3: Summary classification results where *'p'* refers to precision, *'r'* refers to recall, *'domain'* refers to domain-trained embeddings, and *'average'* refers to averaged pre-trained and domain-trained embeddings.

Despite the small amount of data, results in Table 4.3 showed that corpus trained embeddings provide a notable advantage over the larger pre-trained embeddings in

the classifiers performance with achieving F1-score of 0.53 versus F1-score of pre-trained models which is lower than 0.50. Metaembeddings obtained by averaging pre-trained and corpus-trained embeddings did not lead to significant improvements over the corpus-trained embeddings. fastText classifier outperformed GNB model, but only when domain-based embeddings were used (See Table 4.3). A non-verbatim example of a sentence where fastText model, based on corpus-trained embeddings performs better than pre-trained embedding models is: *'The police received information that the subject was selling crack'*. A potential reason for fastText to classify correctly this sentence versus the classifiers using pre-trained embeddings is that the word *'crack'* has the meaning of a *'drug'* in the reports. However, this is not the widely accepted meaning for this word and thus it cannot be interpreted correctly by pre-trained models. A significant problem with using pre-trained models for the safeguarding domain is the high number of polysemous words and abbreviations that appear in the corpus. In such cases, the domain-specific embeddings capture the word meanings better than the pre-trained embeddings (see Table 4.4).

The GNB classifier coupled with pre-trained BERT model outperforms the classifiers based on pre-trained embeddings, however it does not lead to improvements over the domain-based models. Fine-tuning BERT outperforms the other two types of classifiers with micro-F1 of 0.64 and macro-F1 of 0.59 which gives 0.05 improvement over the baseline. The results from the subject-matter experts survey showed that users are satisfied with precision above 0.5 and higher recall, therefore we would deem this results satisfactory. The improvement in the results achieved by fine-tuning BERT indicate the importance of adapting even the more context-aware pre-trained language models to the specific task, especially when the corpus contains domain-specific terminology. Further, the poor performance of classifiers based on pre-trained word models shows the lack of transferability of pre-trained embeddings for a specialised domain such as the safeguarding reports.

Word	Domain-trained Embeddings	Pre-trained Embeddings
crack	heroin, opiates, amphetamines, cocaine, and injecting	cracking, break, shut, knock
battery	affray, acquitted, kindred, gbh, magistrates court	charger, dry-cell, battary, batteries, rechargeable battery
grooming	coercive control, manipulative, sixteen year old, sexual exploitation	self-grooming, shaving, well-groomed, barbering
prism	aid, service, wallich	lens, refract, lense, periscopic

Table 4.4: Comparison between domain-trained embeddings and pre-trained embeddings based on Nearest Neighbour analysis..

4.3.4 Results per Theme

The three best-performing classifiers give similar average results between the dev and test set (see Table 4.5). Further, models tend to return higher results for some themes, especially 'Mental Health Issues' for the test set rather than the dev set. A potential reason for this may be attributed to the fact that the test set has been annotated in a similar manner to the classification models, i.e., independent of the context of the entire documents. The BERT classifier returned high recall and satisfactory precision for the themes 'Contact with Agencies', 'Reflections' and 'Indicative Behaviour' for the dev and test datasets with precision above 0.60 and recall above 0.70. However, the model returns precision below 0.50 for the themes 'Indicative Circumstances' and 'Mental Health Issues' themes. Classification models show better overall performance for the 'Reflections' theme, despite the small amount of labelled data, which is attributed to the more standardised and unified language used across documents.

4.4 Classifiers versus Expert Validators Annotations

In the preceding section we evaluated the performance of the classification approaches against the annotations generated by the creators of the thematic framework, who we refer to as the *expert annotators*. By creating a classifier that uses the annotations

Method	Theme	dev set			test set		
		p	**r**	**F1**	**p**	**r**	**F1**
baseline	Contact with Agencies	.650	.700	.680	.860	.470	.610
	Indicative behaviour	.560	.630	.590	.460	.570	.510
	Indicative circumstances	.330	.640	.440	.520	.510	.510
	Mental Health Issues	.260	.570	.360	.390	.450	.420
	Reflections	.580	.690	.630	.470	.760	.580
	AVERAGE	**.480**	**.650**	**.540**	**.540**	**.550**	**.520**
FT	Contact with Agencies	.550	.750	.630	.790	.740	.760
	Indicative behaviour	.580	.730	.650	.450	.700	.540
	Indicative circumstances	.410	.560	.470	.490	.360	.420
	Mental Health Issues	.260	.420	.330	.350	.310	.330
	Reflections	.580	.640	.610	.510	.640	.560
	AVERAGE	**.480**	**.620**	**.540**	**.520**	**.550**	**.530**
BERT	Contact with Agencies	.620	.820	.710	.840	.580	.690
	Indicative behaviour	.600	.740	.660	.480	.630	.540
	Indicative circumstances	.470	.560	.510	.680	.340	.460
	Mental Health Issues	.310	.510	.390	.470	.460	.460
	Reflections	.590	.760	.670	.510	.820	.630
	AVERAGE	**.520**	**.680**	**.590**	**.600**	**.570**	**.580**

Table 4.5: Results per theme for best performing classifiers where *'AVERAGE'* results are based on macro- measures, *'p'* refers to precision, *'r'* refers to recall, *'FT'* refers to fastText.

generated by expert annotators as a 'ground truth', we aim to produce unified and comparable results across generations that are not susceptible to variations in annotations created by different human annotators interpreting the coding framework. Going further, we judge the predictive power of the models by comparing their performance against the annotations of *expert validators*: independent experts who did not participate in the creation of the thematic annotation framework. We aim to measure the ability of the learned models to conserve the knowledge of the *expert annotators* versus if the task was performed manually by independent experts who were not creators of the framework. In this way, we will be able to judge whether automated approaches are reliable for labeling the reports versus if the task was performed manually by different experts.

The initial coding framework was developed by annotating passages of the documents rather than individual sentences. However, our classifiers are trained with sentences.

In order to fairly judge the predictive power of the models against human coders for annotating sentences and passages of the reports, we performed a study comparing the performance of the classification models versus two independent *expert validators* on both, a sentence and a passage level. For these purposes we used two datasets, one consisting of sentences and one consisting of passages. The *sentence set* consisted of a sample of 100 randomly chosen sentences, while the *passage set* consisted of a 100 passages, each containing three sentences. The *sentence set* was extracted from the dev set while the *passage set* was extracted from the test set (see Table 4.2). We measured the inter-annotator agreement for predicting themes using Cohen's Kappa (see Table 4.6). We also compare the average F1 measure per theme between the expert validators and the best performing classifier (fine-tuned BERT).

	Theme	**Kappa**	**Expert F1**	**BERT F1**
Sentences	Contact with Agencies	.480	.560	.710
	Indicative behaviour	.360	.510	.660
	Indicative circumstances	.320	.390	.480
	Mental Health Issues	.560	.420	.470
	Reflections	.270	.370	.650
	AVERAGE	**.400**	**.450**	**.610**
Passages	Contact with Agencies	.310	.710	.720
	Indicative behaviour	.160	.560	.610
	Indicative circumstances	.380	.540	.580
	Mental Health Issues	.670	.650	.560
	Reflections	.470	.520	.540
	AVERAGE	**.400**	**.600**	**.600**

Table 4.6: Expert validator results based on Cohen's Kappa, and average expert F1, compared to BERT F1, where *'Expert F1'* refers to the average F1 measure between the two expert validators.

The Cohen's Kappa scores showed moderate agreement between the validators with an average score 0.40 on sentence and a passage level. The highest level of agreement is for 'Mental Health Issues' theme. However, the average expert F1 for this theme is surprisingly low. The reason for the discrepancy between the Cohen's Kappa score and the F1 measure is the occurrence of sentences which mention mental health problems such as 'depression'. Such sentences are labelled by the expert validators as 'Mental

Health Issues', however their true label is different because of the surrounding context. Surprisingly, a large portion of these sentences were correctly classified by BERT. The average F1 score for the expert validators significantly improves for passage-level classification with an average F1 = 0.60 in comparison to sentence-level annotations where the average F1 = 0.45 (see Table 4.6). This suggests that humans need more context — i.e., to see the sentences embedded in paragraphs — to classify sentences correctly, compared to deep learning models that can generalize better in these cases with limited context thanks to what they learned from the training set.

4.5 Analysis

We perform two types of analysis. We compare the performance of the classifiers for different length of sentences to observe the classifiers suitability for various sequence lengths. We also measure the effect of the training dataset size on the performance of the models (Section 4.5.1). Additionally, we look at the effect of the number of training instances versus the amount of context provided per instance on the performance of the classifiers (Section 4.5.2).

4.5.1 Effect of Sentence Length and Training Size

Experiments comparing the best-performing classifiers for different sentence length and training set size showed that BERT performed better than the baseline method for any length of sentences. Further, BERT gave higher results than fastText and the baseline for shorter sentences. For long sentences, BERT and fastText had very similar performance with a difference less than 1% (see Figure 4.4). The comparison between the classification models performance for different sizes of training set revealed that deep learning models (i.e., BERT) are highly influenced by the size of the training set in comparison to linear models such as the baseline and fastText (see Figure 4.4). BERT

performed worse than the baseline for the very small training set while fastText gave similar performance to the baseline. However, BERT's performance almost doubled as more sentences were added to the training set while GNB performance was not that heavily influenced by the size of the training data, especially for a training set with more than 1,000 sentences.

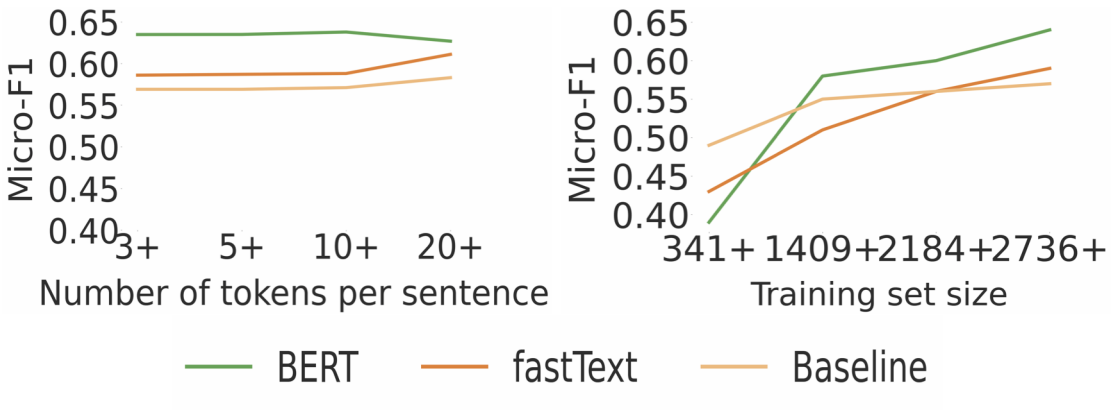

Figure 4.4: Micro-F1 measure per sentence length, i.e., sent with more than 3 tokens, etc. (left) and different train dataset size, i.e., train dataset with up to 341 sentences, etc. (right).

4.5.2 Sentences versus Passages

In this section, we extend the analysis from Section 4.4 by looking at the effect of context versus the number of training instances provided for the classifier models. In this experiment, we gradually increase the length of the training instances in order to judge the importance of the training size versus the context (in terms of passage length). We evaluate the models using sentences and passages where each test passage consisted of three sentences (see Figure 4.5). The evaluation samples for these experiments were extracted from the dev set while the training sentences and passages were extracted from the training set. Results showed that the performance of deep learning models is more influenced by the amount of the training instances rather than the length of

the training passages. Further, models trained on sentence-level with a higher volume of training data giving better results when tested on small paragraphs than classifiers trained on passages but with less training data available. This signifies the importance of higher volume of labelled data for reaching good classifiers performance.

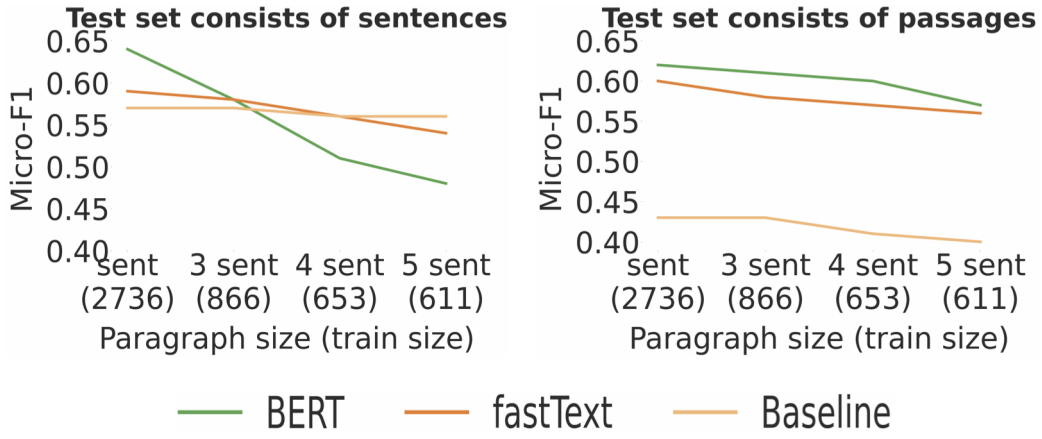

Figure 4.5: Micro-F1 measure per different passage size, where test set consists of sentences (left) and test set consists of passages (right).

4.6 Discussion

4.6.1 Classification Methods

In order to perform a thorough investigation into state-of-the-art techniques, we used multiple approaches for each main stage of the text classification process - feature extraction, feature integration, and classification algorithm as illustrated in Figure 2.5 as part of Section 2.3. Specifically, for feature extraction we used the word embeddings Word2Vec and fastText pre-trained on generic corpora and trained on the specific dataset, as well as the transformer-based language model BERT. We used two approaches for feature integration, including three different types of sentence encoders: InferSent, uSIF, and BERT, each developed using different techniques. Finally, we performed

classification using three classifiers encompassing the main types of classification algorithms, i.e., a simple statistical classifier such as Naive Bayes, a shallow neural network based on embeddings such as fastText classifier, and using fine-tuning technique based on sequential classifier for adapting BERT language model to the classification task.

Developing this classification methodology allowed for in-depth analysis of existing state-of-the-art techniques. Our research extends and complements the work presented by Li et al. [2018] where authors compare word embeddings and sentence encodings for classifying crisis tweets. In contrast, we focus on more specialised domain and classification in low resource settings. Further, we extend the analysis by including transformer-based model and perform more extensive comparison between classification algorithms for different training size and amount of context given.

4.6.2 Findings

Below we present a summary of main findings from this chapter:

1. Simple linear classifier outperforms state-of-the-art language models for very small sized training sets — BERT is outperformed by fastText coupled with domain embeddings and Naive Bayes classifiers for very small volumes of labelled data. However, as the size of the training set increases, BERT outperforms the rest of the models by a significant margin. Further, based on results from expert study where all participants indicated that they are satisfied with precision above 0.500 and high recall, BERT's performance, when the entire training corpus is used, can be considered user satisfactory for the development set as average precision is above 0.500 and recall is above 0.600. However, for the test set, the recall is 0.570.

2. Domain-trained embeddings are more beneficial than large pre-trained embeddings for the classification performance, even when trained on a very small cor-

pus — Feature extraction based on corpus-trained word models coupled with fastText or GNB classifier significantly outperform classifiers build using pre-trained models.

3. Expert validators versus classifiers — This study aimed to judge the predictive abilities of the best performing classifier (BERT) versus expert annotators who did not participate in the original coding of the documents. As original annotations have been carried on passage level while classifiers use sentences, we compared performance for both sentences and passages. Results showed that BERT model performs better than experts on sentence-level and equally well as the human annotators on passage-level. A main finding was that BERT classifies correctly sentences whose label is highly dependent on the surrounding context thanks to what the model learned from the training set.

4. Sentences versus passages — The discrepancy between labels on sentence and passage-level found in Section 4.4 motivated a further research into the effect of the amount of training instances versus the length of training instances (i.e, amount of context) over the classifiers performance. We specifically wanted to identify what is more beneficial for classification models, to be trained on large volume of data consisting of sentences or to be trained on smaller amount of data, consisting of the original coded passages. Results showed that classifiers are more dependent on the size of the training set rather than their length.

The findings from this research are based on the safeguarding reports which are situated in a multi-disciplinary domain involving discussion and language characteristics from various domains such as healthcare, social sciences, and criminology. This makes the findings from this research applicable to healthcare, social sciences, and criminology domains domains. However, to ensure generalisation, we extend on this work by performing quantitative analysis in the next chapter.

4.7 Conclusions

In this chapter, we build on the research of evaluating traditional supervised machine learning approaches by analysing the performance of state-of-the-art classifiers, feature extraction and feature integration techniques. This allowed us to identify classification methods suitable for domain-specific documents such as the safeguarding domain.

Question **RQ2** from the initial hypothesis presented in Section 1.2 has been answered in order to show that state-of-the-art pre-trained language model does lead to better or similar performance to experts who have not participated in the creation of the thematic framework. This shows the potential of contextual language models to perform complex tasks, usually done by domain experts, for specialised corpus. However, analysis also showed that deep learning models performance is highly dependent on the size of the training data in comparison to linear models as BERT's performance is worse than a simple Naive Bayes baseline and fastText for very small training datasets. Further, training word embeddings onto the specific domain, even when the size of the corpus is very small, lead to much higher results in comparison to pre-trained embeddings. This shows the importance of adapting pre-trained models to the specific corpus despite its size.

A main limitation of this research is that analysis are performed for a single domain and classification task. However, validating results require quantitative analysis involving wider range of domains and classification tasks. Further, the current approach is not performing extensive analysis comparing the effect of labelled and unlabelled data over the classifiers performance. The surprising results that simple linear classifiers perform better than state-of-the-art transformer-based models for few-shot classification is the main motivation to explore this topic further for a wider range of datasets in the next chapter.

Chapter 5

Suitability of Text Classification Approaches for Few-shot Settings

The research in the previous chapter presented analysis into state-of-the-art text classification approaches for a small corpus of specialised domain texts. One of the main findings was that simple linear classifiers might be more suitable for few-shot classification than large pre-trained transformer-based models. However, the analysis was limited to a single domain without fully exploring performance of techniques for different sized datasets. In this chapter, we extend on this analysis by investigating the role of task-specific and domain-specific data for the text classification for various domains and classification tasks by looking at how pre-trained and domain-trained models affect the performance. Since it was found in Section 4.5.1 that models perform differently for different length of texts, we differentiate between classification for sentences and documents.

Section 2.2 of the Background chapter presented methods on adapting pre-trained models to the given task or domain. In summary, related studies [Lee et al., 2020, Nguyen et al., 2020, Huang et al., 2019, Alsentzer et al., 2019] investigate whether it is helpful to tailor a pre-trained model to the domain while others [Sun et al., 2019, Chronopoulou et al., 2019, Radford et al., 2018] analyse methods for fine-tuning BERT to a given task. A few research gaps have been identified — lack of extensive analysis for different domains and tasks, lack of comparison between different classification approaches, and lack of consideration of scenarios with limited unlabelled data. A more

extensive research presented by Gururangan et al. [2020], analyse whether it is still helpful to tailor a pre-trained model to the domain of a target task for multiple datasets and tasks. However, this analysis do not focus on few-shot text classification and are limited to exploring performance of a single approach, based on transformer-based model, rather than comparing multiple methods.

In this chapter, we address the aforementioned research gaps by conducting quantitative analysis comparing the light-weight linear classification model fastText, coupled with generic and corpus-specific word embeddings, and the pre-trained language model BERT, trained on generic data and domain-specific data. We also include a simple frequency-based classifier for a baseline. Specifically, we analyze the effect of training size over the performance of the classifiers in settings where such training data is limited, both in few-shot scenarios with a balanced set and keeping the original distributions. We opted for these two models to allow for a comparison with conclusions from previous chapter, and because of their wide application in many NLP tasks. Evaluation of methods has been performed for five domains and six classification tasks.

The rest of this chapter focuses on research question **RQ 3: What are the most efficient approaches for few-shot classification in general and for specialised domains?** from the research questions presented in Section 1.2. Contributions made as part of this research include comparison between state-of-the-art linear classifier and a fine-tuned language model in order to identify their applicability to low resource classification, extensive investigation covering multiple domains into how pre-trained and domain-trained models affect the classification performance.

5.1 Datasets

In this chapter, in addition to the safeguarding reports, we extend analysis for a range of datasets from different domains and nature in order to provide a more conclusive results on suitability of classification methods. The datasets we use are: SemEval 2016

task on Sentiment Analysis [Nakov et al., 2019], SemEval 2018 task on Emoji Prediction [Barbieri et al., 2018], AG News [Zhang et al., 2015], 20 Newsgroups [Lang, 1995] and IMDB reviews [Maas et al., 2011]. One of the reasons for choosing these datasets is that they represent diverse domains (social network, news, reviews) and classification tasks. Further, they allow conducting analysis on sentence- and document- level. We also perform analysis for the safeguarding domain by focusing on multi-class classification. The main features and statistics of each dataset are summarized in Table 5.1.

Dataset	**Task**	**Domain**	**Type**	**Avg length**	**Labels**	**# Train**	**# Test**
SemEval-16 (SA)	Sentiment analysis	Twitter	Sentence	20	3	5,937	20,806
SemEval-18 (EP)	Emoji prediction	Twitter	Sentence	12	20	500,000	49,998
AG News	Topic categorization	News	Sentence	31	4	114,828	5,612
20 Newsgroups	Topic categorization	Newsgroups	Document	285	20	11,231	6,728
IMDB	Polarity detection	Movie reviews	Document	231	2	28,000	23,041
Safeguarding reports	Theme detection	Safeguarding	Sentence	18	5	2,494	284

Table 5.1: Overview of the classification datasets used in our experiments, where *'# Train'* indicates the number of training instances in the given dataset split and *'# Test'* indicates the number of test instances in the given dataset split.

5.1.1 Social Media Datasets

We use two Twitter collections gathered for the conduct of various NLP challenges and competitions.

SemEval-2016 task 4 (SE-16) The SemEval-2016 dataset have been initially collected and used for the 'Sentiment Analysis in Twitter Task' challenge. The challenge consists of four subtasks. We focus on subtask A, where the goal is for a given tweet to predict whether it is of positive, negative, or neutral sentiment [Nakov et al., 2019].

SemEval-2018 task 2 (SE-18) The SemEval-2018 dataset has been used as part of a competition on emoji prediction. Given a text message including an emoji, the emoji prediction task consists of predicting that emoji by relying exclusively on the textual content of that message. In particular, task focuses on the one emoji occurring inside tweets. The 20 most frequent emojis are considered for the task [Barbieri et al., 2018].

5.1.2 Newsgroups and News

We use two news-related datasets: 20 newsgroups collection [Lang, 1995] and the AG news collection [Zhang et al., 2015].

AG News The AG news collection [Zhang et al., 2015] consists of news articles that have been gathered from more than 2000 news sources by 'ComeToMyHead' in more than 1 year of activity. 'ComeToMyHead' is an academic news search engine which has been running since July, 2004 [1]. The dataset is provided by the academic community for research purposes in data mining. The classification task consists of topic categorization and can be represented as a multi-class problem where each class presents news category, i.e., sports, politics, etc. Each news instance contains a title and a description. We combine both for our analysis.

20 Newsgroups The 20 newsgroups collection [Lang, 1995] is a popular data set for experiments in machine learning. The data is organized into 20 different newsgroups, each corresponding to a different news topic such as computer systems, religion, politics. [2]

[1] AG's corpus of news articles: `https://cloud.google.com/apis/`

[2] The 20 Newsgroups corpus: `http://qwone.com/~jason/20Newsgroups/`

5.1.3 Movie Reviews

We use the Internet Movie Database (IMDB) movie reviews dataset [Maas et al., 2011] for some of our analysis. The collection consists of movie reviews used for binary sentiment classification, i.e., identifying positive and negative movie reviews.

5.1.4 Safeguarding Domain

In contrast to the previous two chapters, here we convert the multi-label classification task for the safeguarding reports to multi-class problem by removing instances that were labelled with more than one class in the original datasets. One of the reasons for introducing this change is to allow for more clarity and simplicity when we compare results per datasets for balanced datasets. Further, this help comparison between results in this chapter and results, presented in Chapter 6. However, we explain analysis choices and dataset adjustments relevant to Chapter 6 in Section 6.1. We present examples of each dataset in Table 5.2.

Dataset	Example	Label
SemEval-16 (SA)	After attempting a reinstall, it still bricks, says, 'Windows cannot finish installing,' or somesuch. @Microsoft may have cost me $600.	negative
SemEval-18 (EP)	My view tonight...Is this healthy since it has lettuce on it @ Roosters 2	_face_with_tears_of_joy_
AG News	Stocks Climb Despite Rising Oil Prices. Stocks turned higher Tuesday, sending the Dow Jones industrial average back above the 10,000 mark, as investors shrugged off rising energy prices and focused instead on good corporate news.	Business
20 Newsgroups	Geez. Everyone comes up with Clark, Williams, Thompson. These guys were all up in 1987. That's ancient history. So in the last 6 years, noone, right? Beck doesn't count. I said 2 solid years. BTW, Manwaring lead the ML last season in throwing out baserunners. He is an excellent defensive catcher. I agree that his offensive skills are limited but he does seem to be improving on them. Let's see what he does w/o the help of a pitchout every other pitch. As I remember, even Bob Brenly had a good throwout percentage under Roger Craig, who loved to sacrifice the count for runners being thrown out. Of course, he suffered from 3 ball 1 strike homers a lot too. I am not a big fan of Manwaring.	rec.sport.hockey
IMDB	I went and saw this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that. The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great, and I suggest that you go see it before you judge.	positive
Safeguarding reports	It is worth noting that there are potential issues around over-reliance on email communication between agencies throughout this case, but in particular between Probation and Housing.	Reflection

Table 5.2: Datasets examples.

5.2 Experimental Setting

5.2.1 Comparison Systems

We compare fastText (FT) classifier and BERT because results in Chapter 5 showed that these are the best performing classifiers for the safeguarding domain. Further, analysis presented in Section 4.5 showed that fastText performs better for certain scenarios with very limited data. Additionally, BERT gives a state-of-the-art performance for various NLP tasks. In this chapter, we take these analysis further. We also include a simple baseline based on frequency-based features and a suite of classification algorithms available in the Scikit-Learn library [Pedregosa et al., 2011], namely GNB, Logistic Regression (LG) and SVM. Of the three, the best results for the majority of the six datasets were achieved using Logistic Regression. Therefore, we include LG model as a baseline for the experiments in this chapter.

5.2.2 Training

Similarly to analysis performed in Chapter 4, as pre-trained word embeddings we downloaded 300-dimensional fastText embeddings trained on Common Crawl. In order to learn domain-specific word embedding models we used the corresponding training sets for each dataset, except for the Twitter datasets where we leveraged an existing collection of unlabelled tweets from October 2015 to July 2018 to train 300-dimensional fastText embeddings [Camacho Collados et al., 2020]. Word embeddings are then fed as input to a fastText classifier where we used default parameters and softmax as the loss function. As for BERT, we fine-tune it for the classification task using a sequence classifier, a learning rate of 2e-5 and 4 epochs. In particular, we made use of BERT's Hugging Face default transformers implementation for classifying sentences [Wolf et al., 2019] and the hierarchical principles described by Pappagari et al. [2019] for pre-processing long texts before feeding them to BERT. We used the

generic base uncased pre-trained BERT model and BERT-Twitter[3], both from Hugging Face Wolf et al. [2019].

5.2.3 Evaluation Metrics

We report results based on standard micro and macro averaged F1 [Yang, 1999] as we have done in previous chapters. In our setting, since system provides outputs for all instances, micro-averaged F1 is equivalent to accuracy.

5.3 Analysis

We perform two main types of analysis. First, we look at the effect of training size over the classifier's performance by randomly sampling different sized subsets from the original labelled datasets (Section 5.3.1). Then, we perform a few-shot experiment where we compare classifier's performance on different sizes of balanced subsets of the training data (Section 5.3.3).

5.3.1 Effect of Training Set Size

Table 5.3 and Figure 5.3.1 show the results with different sizes of training data randomly extracted from the training set. Surprisingly, classification models based on corpus-trained embeddings achieve higher performance with less labelled data compared to the classifier based on pre-trained contextualised models. However, for cases with more than 5,000 training samples, the performance of fine-tuned BERT significantly outperforms fastText corpus-based classifier, especially when domain-trained BERT model (i.e., BERT (Twitter)) is used. Further to that, the fine-tuned model performance improves at a higher rate than the classifier based on corpus-trained embed-

[3]BERT-Twitter model: `https://huggingface.co/ssun32/bert_twitter_turkle`

	Model	Micro-F1						Macro-F1					
		200	**500**	**1000**	**2000**	**5000**	**ALL**	**200**	**500**	**1000**	**2000**	**5000**	**ALL**
SE-16 (SA)	Baseline	.399	.405	.430	.447	.476	.483	.360	.390	.410	.430	.430	.460
	FT (general)	.423	.453	.460	.478	.528	.530	.393	.446	.455	.470	.480	.490
	FT (domain)	**.463**	**.487**	.490	.497	.542	.560	**.446**	**.481**	.484	.490	.500	.520
	BERT (general)	.381	.438	**.547**	**.546**	.600	**.611**	.300	.400	**.530**	**.540**	.580	.600
	BERT (Twitter)	.422	.461	.527	.544	**.603**	**.611**	.330	.450	.520	**.540**	**.590**	**.610**
SE-18 (EP)	Baseline	.108	.116	.133	.139	.154	.201	.100	.110	.130	.140	.150	.190
	FT (general)	.109	.120	.130	.136	.194	.258	.084	.109	.125	.130	.150	.220
	FT (domain)	**.149**	**.153**	**.160**	**.166**	.219	.262	**.108**	**.135**	**.151**	**.160**	**.180**	.220
	BERT (general)	.040	.097	.106	.120	.261	.380	.010	.030	.050	.070	.120	.280
	BERT (Twitter)	.082	.102	.134	.127	**.291**	**.400**	.030	.060	.080	.110	.150	**.380**
AG News	Baseline	.665	.788	.812	.840	.854	.883	.620	.750	.780	.800	.820	.860
	FT (general)	.609	.761	.799	.836	.885	.901	.548	.720	.758	.800	.860	.877
	FT (domain)	**.857**	**.886**	.884	.889	.902	.905	**.831**	**.858**	.857	.860	.880	.881
	BERT (general)	.600	.856	**.910**	**.910**	**.922**	**.954**	.390	.830	**.880**	**.880**	**.900**	**.940**
20 Newsgroups	Baseline	.323	.401	.453	.495	.512	.534	.310	.390	.450	.490	.510	.530
	FT (general)	.311	.409	.510	.567	.620	.633	.275	.398	.490	.560	.610	.624
	FT (domain)	**.458**	**.533**	**.583**	**.621**	.636	.645	**.440**	**.520**	**.573**	**.610**	.630	.630
	BERT (general)	.079	.192	.485	.637	**.700**	**.714**	.040	.110	.420	.590	**.670**	**.700**
IMDB	Baseline	.770	.787	.810	.835	.843	.857	.770	.787	.810	.835	.841	.857
	FT (general)	.750	.770	.811	.845	.859	.878	.750	.771	.811	.845	.859	.878
	FT (domain)	**.814**	**.815**	**.836**	**.862**	**.871**	.879	**.814**	**.815**	**.836**	**.862**	**.871**	.879
	BERT (general)	.543	.783	.834	.850	.850	**.881**	.543	.783	.834	.850	.850	**.881**
Safeguard reports	Baseline	.357	.431	.431	.452	-	.484	.264	.336	.363	.389	-	.404
	FT (general)	.420	.421	.484	.452	-	.494	.301	.324	.355	.346	-	.364
	FT (domain)	**.421**	**.463**	**.526**	**.526**	-	**.505**	**.316**	**.351**	**.391**	**.474**	-	**.477**
	BERT (general)	.231	.326	.431	.473	-	.494	.124	.217	.304	.353	-	.370
AVERAGE	Baseline	.437	.488	.488	.535	.568	.574	.404	.460	.491	.491	.550	.550
	FT (general)	.437	.489	.532	.552	.617	.616	.392	.461	.494	.525	.592	.576
	FT (domain)	**.527**	**.556**	**.579**	**.589**	.634	.626	**.492**	**.527**	**.548**	**.576**	.612	.601
	BERT (general)	.312	.449	.552	**.589**	**.667**	**.672**	.234	.395	.503	.547	**.624**	**.628**

Table 5.3: Results by training size: 200, 500, 1000, 2000, 5000 instances and entire training set (ALL), where each subset is extracted from the larger subset.

dings for training sets with more than 2,000 instances. For instance, for the SE-18 dataset, fastText with domain embeddings improves 0.112 micro-F1 points when the entire dataset is used with respect to using only 200 instances, while BERT-Twitter provides a 0.360 absolute improvement. In contrast, fastText with pre-trained embed-

dings performs similarly to the baseline. This shows the advantage for pre-trained models to be fine-tuned to the given domain and task.

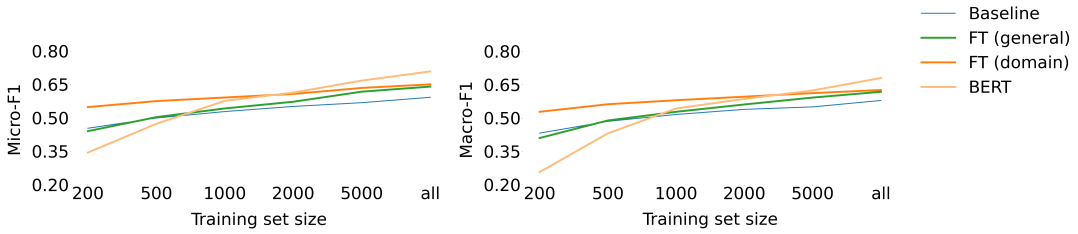

Figure 5.1: Experiments with random data distribution, where Micro-F1 results (left), Macro-F1 results (right).

5.3.2 Sentences versus Documents

In order to avoid confounds such as the type of input data in each of the experiments, we filter the results by sentences and documents (see Table 5.1 for the actual split of datasets in each category). Figure 5.2 shows the results for this experiment. As can be observed, training set size affects similarly for both types of input, with BERT being especially sensitive to the training data size. However, whilst we already showed that BERT is highly sensitive to training set size fastText seems far more reliant on context with fasText results at a document level and a small training set comparing favourably to how it performs at a sentence level even when the entire dataset is used for training.

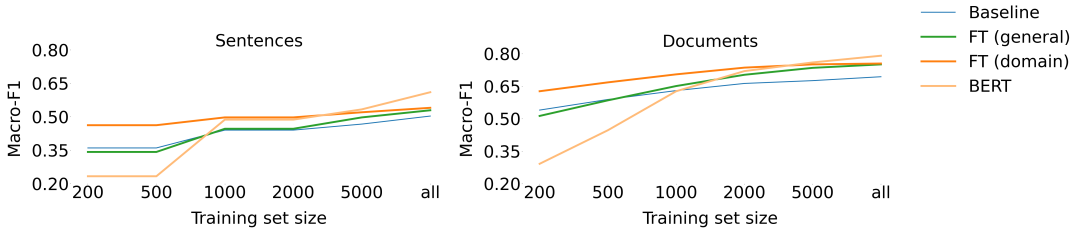

Figure 5.2: Macro-F1 results with randomly sampled training data split by type: sentence or document.

5.3.3 Few-shot Experiment

A few-shot comparison between the performance of classifiers based on balanced data is shown in Table 5.4. The full balanced set is built by removing random training instances based on the least frequent label's number of instances. We balance the dataset for a few shot experiments to ensure the occurrence of instances for all labels within the training set even for datasets with 20 labels when 5-shot and 10-shot experiments are performed. Further, we look at the effect of balanced training data over the classifiers performance. The results show that balancing the dataset lead to improvements in the classification performance with limited training data, especially for BERT. For example, using a subset of 1,000 training instances for 20 Newsgroups corpus, the macro-F1 for random sampled data is 0.420 while the macro-F1 for balanced data (i.e., 50 instances per label) is 0.556.

	2 instances per label				**5 instances per label**				**10 instances per label**			
	FT(gen)	FT(dom)	BERT(gen)	BERT(T)	FT(gen)	FT(dom)	BERT(gen)	BERT(T)	FT(gen)	FT(dom)	BERT(gen)	BERT(T)
SE-16(SA)	.330	**.390**	.230	.320	.370	**.390**	.240	.370	.352	**.384**	.200	.370
SE-18(EP)	.050	**.060**	.020	.030	.080	**.100**	.020	.040	.090	**.110**	.020	.050
AG News	.390	**.700**	.130	-	.520	**.810**	.200	-	.643	**.815**	.410	-
20 News	.090	**.200**	.010	-	.230	**.430**	.030	-	.294	**.473**	.030	-
IMDB	.411	**.556**	.492	-	.500	**.643**	.547	-	.414	**.567**	.492	-
Safeguard	.181	**.251**	.070	-	.238	**.296**	.070	-	.236	**.246**	.150	-
AVERAGE	.242	**.359**	.159	-	.323	**.445**	.185	-	.339	**.431**	.217	-
	20 instances per label				**50 instances per label**				**Full balanced dataset**			
	FT(gen)	FT(dom)	BERT(gen)	BERT(T)	FT(gen)	FT(dom)	BERT(gen)	BERT(T)	FT(gen)	FT(dom)	BERT(gen)	BERT(T)
SE-16(SA)	.356	**.406**	.320	.370	.416	**.466**	.340	.420	.510	.530	**.610**	.570
SE-18(EP)	.096	**.126**	.020	.070	.121	**.144**	.060	.100	.160	.170	.200	**.280**
AG News	.686	**.838**	.680	-	.752	**.845**	.740	-	.860	.880	**.940**	-
20 News	.406	**.537**	.090	-	.499	**.568**	.500	-	.620	.640	**.680**	-
IMDB	.496	**.641**	.504	-	.660	**.707**	.556	-	.870	.880	**.882**	-
Safeguard	.180	**.260**	.220	-	.356	**.376**	.160	-	.432	**.448**	.360	-
AVERAGE	.370	**.468**	.306	-	.467	**518**	.393	-	.578	.589	**.612**	-

Table 5.4: Few-shot Macro-F1 classification results, where *'gen'* refers to general and *'dom'* refers to domain, and *'BERT(T)'* refers to BERT-Twitter model, trained using Twitter data.

Similarly to the experiments with randomized data samples, fastText based on corpus-trained embeddings is the best performing classification model for very small amounts

of balanced labelled data (see Figures 5.3 and 5.4). However, as the amount of training data increases, BERT model outperforms fastText on average by 0.0442%. As in the previous experiment, the classification model based on pre-trained embeddings perform poorly compared to the corpus-trained embeddings and models fine-tuned to the task. Further, BERT (Twitter) leads to significant improvements over BERT when only 10 instances per label are used (i.e., for SE-16, BERT (Twitter) has macro-F1 = 0.370, similar to domain-based fastText with macro-F1 = 0.384 versus base BERT with macro-F1 = 0.200).

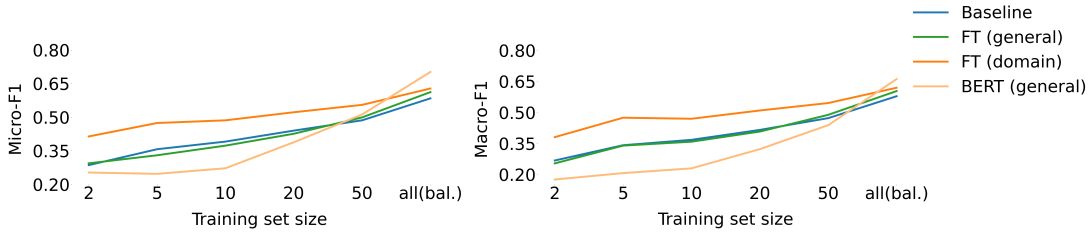

Figure 5.3: Experiments with balanced data, where Micro-F1 results (left) and Macro-F1 results (right).

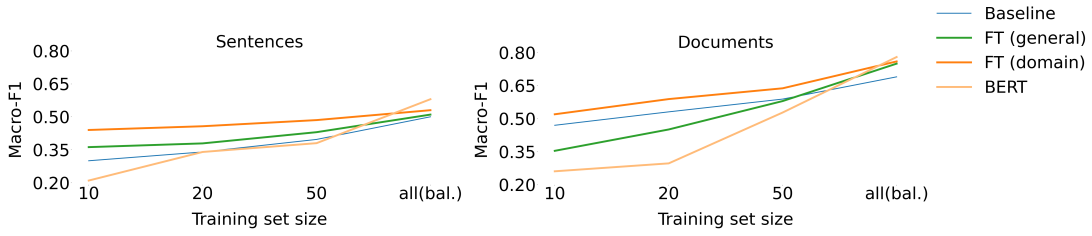

Figure 5.4: Macro-F1 results with balanced training data split by type: sentence or document.

5.3.4 Word embeddings: Coverage and Nearest Neighbours

A comparison between the number of OOV words for the test datasets between the generic and the domain-trained models (see Table 5.5) shows that the domain-trained

model vocabularies have a larger number of OOV words for the test set than the bigger more generic models except SemEval-18 dataset (Emoji Prediction).

Dataset	# Tokens	# OOV domain	# OOV general
SemEval-16 (SA)	30,467	12,668 (41.6%)	10,558 (34.7%)
SemEval-18 (EP)	66,294	36,846 (55.5%)	37,335 (56.3%)
AG News	23,024	7,766 (33.7%)	4,712 (20.5%)
20 Newsgroups	79,343	39,056 (49.2%)	27,970 (35.2%)
IMDB	105,240	54,395 (51.6%)	36,223 (34.4%)
Safeguard	12,780	633 (49.5%)	25 (1.95%)

Table 5.5: Number of tokens and OOV tokens for the domain-specific (*'#OOV domain'*) and general-domain word embeddings (*'#OOV general'*) models per test set.

However, the generic domain embeddings tend to fail to represent the meaning of more domain-specific words, which may explain their lower performance. This is confirmed by the nearest neighbour analysis (see Table 5.6) which showed that the generic domain embeddings do not provide accurate representations of more technical words such as *'Windows'* and *'Sun'*. In the IMDB reviews, words such as *'Toothless'*, used within a very specific context are also not correctly represented by the generic model. Moreover, tweets are rich with abbreviations which have domain-specific meaning such as *'SF'* referring to *'San Francisco'*.

5.4 Discussion

5.4.1 Quantitative Analysis

The quantitative analysis include a wide coverage of five domains and six classification tasks. This analysis is similar to the work presented by [Gururangan et al., 2020] where authors analyse different methods for pre-training and fine-tuning roBERTa model using four domains and eight tasks. In contrast, our research is entirely focused on the

Vocabulary	**Token**	**FT (domain)**	**FT (generic)**
Twitter	SF	San Francisco	SciFi
	killing	killin'em	murdering, slaughtering
	arsenal	arsenal fc	armoury
AG News	Sun	Microsystems	Sunlight
	Apple	iTunes	Pear
	capsule	spacecraft	pill
20 Newsgroups	Windows	X Windows	Window, Doors
	DOS	DOS 6, DR-DOS	don't
	backdoor	eavesdrop, algorithm	back-door,loophole
IMDB	Toothless	Worthless	Toothlessness
	slow-pacingly	boringly	fast-pacing
	twist	plot twist	spin
Safeguard	idva	forensic	suaminya
	order	sentenced	ordering
	crown	robbery, court	crowns, jewel

Table 5.6: Examples of words and their nearest neighbour according to the generic (*'FT (generic)'*) and domain-specific word embedding models (*'FT (domain)'*).

classification task considering a wide range of low resource settings. Further, we perform comparisons between different classification algorithms rather than analysing a single model. This makes the analysis in this chapter, the most extensive work conducted to date on suitability of state-of-the-art neural models for low resource classification. We further build on previous research [Gururangan et al., 2020, Sun et al., 2019, Chronopoulou et al., 2019, Radford et al., 2018] focusing only on language models, by performing a comparison between two classification approaches, based on fastText and BERT. We also include a frequency-based Logistic Regression classifier as a baseline. Further, we perform experiments considering generic and domain-trained embedding and language models. To ensure coverage of different low resource classification scenarios, we performed two types of analysis: 1) random sample distribution with training instances ranging from 200 to the entire training set and 2) few shot classification with

training instances ranging from 2 instances per label to the entire balanced training set. Further, we look at whether classification performance differ for sentences versus documents.

5.4.2 Findings

The main findings from this chapter are summarised below:

1. For small datasets, especially for collections with less than 2000 training instances or for few-shot scenarios with less than 50 instances per label, it is more beneficial to use domain-trained word embeddings coupled with less resource consuming linear classifiers such as fastText rather than using pre-trained or domain-trained language model such as BERT.

2. Language models improve their performance on a higher rate than linear classifiers coupled with embeddings as more training data is used — Pre-trained and especially domain-trained BERT outperforms fastText classifier for larger training samples and improves at a higher rate.

3. Balancing datasets does lead to classifiers improvements versus random distribution especially for BERT classifier.

4. fastText is more reliant on context than BERT — fastText performs better for classifying documents than sentences while BERT performance is much more sensitive to the training size than the context provided.

5.4.3 Limitations

The analysis in this chapter can be extended to more classification tasks and different models, e.g., larger language models such as RoBERTa [Liu et al., 2019]. Further,

work could be improved by exploring more approaches for adapting models to the domain and task such as using continuous pre-training on the domain and task data. It can also be beneficial to extend analysis on the role of unlabelled data for text classification by using meta-embeddings such as concatenating BERT and domain-trained fastText embeddings or by including more domain-trained language models. Despite this limitations, our analysis were conclusive in that simpler less resource consuming methods, based on domain-trained embeddings are more suitable for few-shot classification than pre-trained or domain-adapted state-of-the-art language models.

5.5 Conclusions

In this chapter, we expanded analysis on suitability of state-of-the-art classification methods for limited sized datasets by conducting a quantitative study looking at the effect of both, training and unlabelled domain-specific data over supervised text classification. We compared linear and transformer-based language models in few-shot scenarios with a balanced set and by randomly sampling different sized subsets from the original labelled datasets. The research helped identify what feature extraction and classification approaches are suitable for limited training datasets.

Question **RQ3** has been answered in order to show that in settings with small training data, a simple linear classifier such as fastText coupled with domain-specific word embeddings appear to be more robust than a more data-consuming model such as BERT, even when BERT is pre-trained on domain-relevant data. However, the same classifier with generic pre-trained word embeddings does not perform consistently better than a traditional frequency-based baseline. BERT, pre-trained on domain-specific data (i.e., Twitter) leads to improvements over generic BERT, especially for few-shot experiments.

In this chapter and the previous chapter, we have focused on analysing existing classification methods and their suitability for small specialised corpora. In the rest of the

thesis, we will build on this research by developing methodologies for improving the performance of these classification methods.

Chapter 6

Text Generation-based Data Augmentation Techniques for Few-shot Text Classification

The quantitative analysis from the previous chapter gave an insight on the type of classification strategies suitable for domains with limited volume of data. In this chapter, we build on these findings by introducing a data augmentation method that helps improve the performance of these classification strategies, specifically for few-shot scenarios. We opted for using text generation techniques as these techniques provide more diversity and theoretically much larger volume of artificial samples in comparison to WR-based and SR-based DA techniques (outlined in Section 2.3.8).

Despite the clear advantages of using TG-based DA methods such as fast, low-cost and less labour intensive generation of additional data compared to manual annotation, a main problem with such approaches is the possibility of generating noise which decreases the performance of classification models rather than improving it [Yang et al., 2020]. As outlined in Section 2.3.8, this noise distribution problem has been ignored in previous research, where approaches are focused on the creation of label-preservation techniques for the generated synthetic data samples and comparing different TG techniques [Anaby-Tavor et al., 2020, Wang and Lillis, 2019, Zhang et al., 2020a, Kumar et al., 2020]. This is also the main motivation for the research presented in this chapter. Our aim is to improve the quality of generated artificial instances and thus

improve classifiers, by developing seed selection strategies for choosing class representative seed samples used for producing the artificial data. Additionally, we analyse how different approaches of fine-tuning GPT-2 model affect the quality of generated data and consequently the classification performance. Specifically, we identify which fine-tuning method leads to generating higher quality data — fine-tuning on a smaller but label-relevant dataset or fine-tuning on larger unlabelled set.

The proposed methodology complements the research presented by Yang et al. [2020] where authors describe an approach based on the use of influence functions and heuristics for selecting the most diverse and informative artificial samples from the already generated artificial dataset. Instead, we focus on seed selection strategies for selecting the most informative samples from the original data. We hypothesise that focusing on selecting the most class representative samples from the original data in the first place can already lead to important improvements and has an important efficiency advantage, as it prevents an unnecessary waste of resources and time of generating unused documents, especially considering how resource expensive generative language models are. Further, Yang et al. [2020] focus on a single domain, i.e., common sense reasoning without considering few-shot settings, while we explore performance of strategies for three domains. The seed selection strategies presented in this chapter have not been used in previous research and represent a novel approach for improving the performance of TG-based DA methods.

In the rest of the chapter, the final question **RQ 4: Can text classification performance be improved through the use of data augmentation techniques based on text generation and seed selection strategies in few-shot settings in general and for specialised domains?** in Section 1.2 is addressed in order to show whether seed selection strategies can help generate higher quality artificial data and thus help improve classification. Contributions made as part of this research include:

- A comparison between four seed selection strategies.
- A comparison of two methods for adapting the text generative models to the

classification task.

- An extensive evaluation of strategies for three domains and four baselines.

6.1 Data Augmentation Methodology

The methodology consists of devising seed selection strategies that help select the most class representative training instances. These seeds can be used by text generation techniques to produce higher quality additional training data and subsequently improve the classifiers performance. Additionally, we experiment with two fine-tuned GPT models in order to identify optimal techniques for adapting GPT-2 model for DA techniques for classification. Many specialised domains, such as the safeguarding reports, are associated with hierarchical class structure which is usually created by a domain expert. Such hierarchy can be a valuable source of information, which can be used as a guidance to improve the prediction of the overall classes. In order to test this hypothesis for a wider range of domains, in addition to the safeguarding reports, we use two additional datasets with a hierarchically organised labels. We explain the seed selection strategies in depth in Section 6.1.1. Further, our analysis focus on few-shot classification because of the high need for approaches which can perform well for only a handful of training instances especially in specialised domains where experts are sparse and data access is limited, as explained in Section 2.3.5. However, our methodology can be easily extended for classification problems with more labelled data and it can also be used to generate more artificial training data, which we plan on analysing in future.

We limit the analysis to multi-class problems to provide more clarity and transparency into the sample generation process with a common evaluation setting, similarly to analysis presented in Chapter 5. Also, focusing on samples with a single label can further help generate stronger class representatives and thus help both multi-class and multi-label classification, which we plan to analyse in future work. The DA methodology consists of three main steps (see Figure 6.1). In the first step, *Seed Selection*, we select

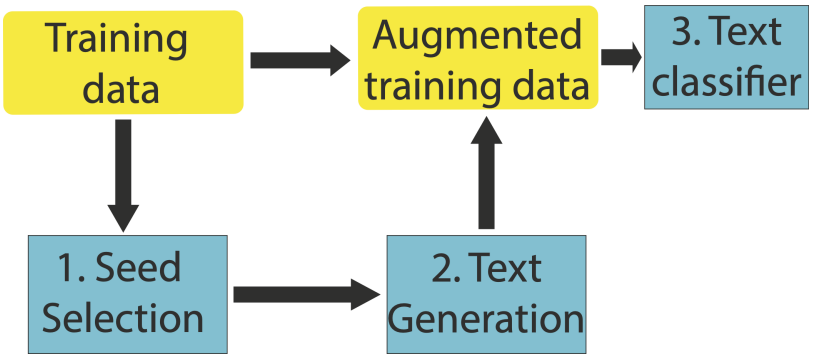

Figure 6.1: Overview of the methodology

samples (i.e., *seeds*) from the original labelled data, used for generating new training instances. In the second step, *Text Generation*, we generate additional artificial training data using text generative model. Finally, the *Augmented training data* is used in combination with the original data to train a *Text classifier*.

6.1.1 Seed Selection Strategies

We implement four seed selection strategies, which are Random Seed Selection, Maximum nouns-guided Seed Selection, Subclass-guided Seed Selection, and Expert-guided Seed Selection. We chose these seed selection strategies because they exploit characteristics associated with specialised domains such as high number of terms, annotation performed by experts, and hierarchical class structure (common for social science and medical domains which require thematic analysis). Further, these characteristics can be used to identify any domain-specific language. The strategies are described below.

Random Seed Selection We simply select a fixed number of instances per label in a random manner. We use random selection to evaluate whether the rest of the seed selection strategies lead to improvements in classification.

Maximum Nouns-guided Seed Selection As highlighted in Section 3.1, the safeguarding reports are rich of domain-specific terminology and thus we believe that noun-rich instances might be more indicative for the classes compared to the other training samples. To test the truthfulness of this hypothesis, we use *Maximum nouns-guided Selection* strategy where we select the seeds with the maximum occurrence of nouns. We identify nouns within data using NLTK.

Subclass-guided Seed Selection In this strategy, we leverage the human-generated classification hierarchy of a dataset in order to improve the classification of the top classes. Specifically, we select a roughly balanced number of seeds from each subclass belonging to a given label. In this way, we diversify the vocabulary for each overall class by ensuring the equal participation of representative samples from even the most underrepresented subclasses. For the purpose of the analysis we use only the first level of subclasses.

Expert-guided Seed Selection The highly specialised nature of the safeguarding reports and the fact that the manual annotation of documents need to be performed by experts show that identifying theme representative samples might require more implicit knowledge that is hard to be captured by statistical approaches and require an expert. Therefore, in the *Expert-guided Seed Selection* strategy, we conducted a study asking experts to select the class representative samples from the original training data. The chosen seeds are used to generate additional training data. We performed this strategy only for the safeguarding reports because the other two datasets do not require expert annotators. We explain further this method, including description of experiments and results in Section 6.4.

6.1.2 Text Generation

We generate artificial data using the generative pre-trained model, GPT-2. We use GPT-2 model as it gives a state-of-the-art performance for many text generation tasks and also have been designed with the objective to fit scenarios with few-shot and even zero-shot settings, as described in Section 2.3.6. We use two methods for fine-tuning the GPT-2 model — we fine-tune the model on the entire dataset and we also fine-tune a specific GPT-2 model for each given class to ensure label-preservation for the generated sequences. We also perform experiments using a pre-trained GPT-2 model. We compare this three models in order to assess the need of fine-tuning and the use of additional methods for label-preservation when using TG-based DA for classification tasks. These models are then leveraged to generate new documents given a labelled instance.

Fine-tuning GPT-2 per Label Technique. In Section 2.3.8 we outlined two main methods for label preservation of generated samples. The first approach, using a classifier to re-label artificial sequences, requires either a large training corpus to ensure high performance of the classifier in first place or the generation of large volume of artificial data to ensure that a substantial amount of these will not be filtered because of a low threshold [Anaby-Tavor et al., 2020]. The other, more widely accepted approach, is prepending the class labels to text sequences during fine-tuning of the Transformer-based model [Wang and Lillis, 2019, Zhang et al., 2020a, Kumar et al., 2020]. Such an approach cannot ensure label-preservation for all generated sequences. However, our priority is to allow a fair comparison for seed selection approaches without introducing additional noise. Therefore, we consider a simple technique based on fine-tuning a model per label more suitable for performing our analysis.

6.1.3 Classification

In the final step, *Classification*, we use the *Augmented training data* to train a fastText classifier coupled with domain-trained fastText word embeddings. The reason to use fastText is because in Chapter 5, we showed that fastText classifier combined with domain-trained embeddings outperforms state-of-the-art models such as BERT for few-shot scenarios. As explained in Section 2.1.4 and confirmed by the analysis performed in Chapter 4, fastText embeddings tend to deal with OOV words better than Glove and Word2Vec approaches. Also, fastText embeddings are the default using the fastText classifier. Therefore, we use fastText embeddings for performing analysis in this chapter.

6.2 Datasets

Similarly to the previous chapter, in addition to the safeguarding reports, we use two additional datasets which are publicly available, have been used in classification tasks and allow evaluation of the data augmentation approach on a wider scale. Specifically, we use the 20 Newsgroups [Lang, 1995], already introduced in Chapter 5, and Toxic comments [Hosseini et al., 2017]. We perform prediction for the top classes of the datasets, however, as mentioned in Section 6.1, we use the sub-classes to select seed instances. As mentioned earlier in the chapter, for simplicity and setting unification, we convert the multi-label classification tasks of the datasets to multi-class problems, removing the instances that were labelled with more than one class in the original datasets. The main features and statistics of each dataset are summarized in Table 6.1[1].

Classification Hierarchy of the Datasets The *Safeguarding reports* (see Figure 6.4 for a reference of number of sentences and passages for the most underrepresented

[1]We have included statistics and classification results of unmodified datasets in Appendix A, Section 7.5

Dataset	**Task**	**Domain**	**Av len**	**Class**	**Subclass**	**# Test**
20 Newsgroups	Topic cat	Newsgroups	285	6	20	6,728
Toxic comments	Toxic pred	Wikipedia	46	2	5	63,978
Safeguarding reports	Theme pred	Social	45	5	34	284

Table 6.1: Overview of the text classification datasets, where *'#Test'* indicates the number of instances in the test set and average length (*'Av len'*) is measured as the average number of tokens per instance.

theme 'Mental Health Issues') consists of 5 overall classes and 34 subclasses while the *20 Newsgroups* dataset (see Figure 6.2 for a reference of number of training instances per class and subclass) there are 20 subclasses split between 6 overall classes. The collection of the *Toxic comments* dataset (see Figure 6.3 for a reference of number of training instances per classes and subclasses) is obtained from Wikipedia and it is the result from the collaboration between Google and Jigsaw for creating a machine learning-based system for automatically detecting online insults, harassment, and abusive speech [Hosseini et al., 2017]. The classification task consists of predicting different level of toxicity in user comments on Wikipedia. Originally, there are 8 classes: 'non-toxic', 'toxic', 'severe toxic', 'obscene', 'threat', and 'identity hate' where the toxic-related labels can overlap. For providing a hierarchical classification structure, we combine all toxicity-related labels under the label 'toxic'. In this way, we create a class hierarchy with two overall classes - 'toxic' and 'non-toxic' where the 'toxic' class is overarching 6 subclasses. An overview of the classes and the subclasses of the three datasets is given in Table 6.2.

Filtering Training Data We focus on a few-shot scenarios where the dataset is balanced. We start experiments with 5 and 10 instances per label, extracted randomly from the original data ('base' instances), with at least one instance per subclass. Then, we add 5, 10, and 20 artificially generated instances to the 'base' instances ('add' instances) in order to evaluate the effect of methods over different sized training data

Dataset	Class	Sub-classes
20 Newsgroups	computers	comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x
	recreational activities	rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey
	science	sci.crypt, sci.electronics, sci.med, sci.space
	forsale	misc.forsale
	politics	talk.politics.misc, talk.politics.guns, talk.politics.mideast
	religion	talk.religion.misc, alt.atheism, soc.religion.christian
Toxic comments	non-toxic	non-toxic
	toxic	mild toxic, severe toxic, obscene,threat, insult,identity hate
Safeguarding reports	Contact with Agencies	Health Practitioners, Contact with Third sector orgs, Educational Institutions, Contact with Social Care, Police Contact, Contact with councils or LAs
	Indicative Behaviour	Lying, Offending, Serious Threats to Life, Weapons, Emotional Abuse, Domestic Violence, Substance Misuse, Alcohol Misuse, Harassment, Self Inflicted Harm, Stalking, Controlling Behaviour, Aggression
	Indicative Circumstances	Bereavement,NFA, Homelessness or Constantly changing Address, Family Structure, Child Safeguarding, Relationship Breakdown, Debt or Financial Exploitation, Sex Work, Relationship with Children, Quality of Relationship
	Mental Health Issues	Children, Victim, Perpetrator, Suicidal Ideation
	Reflections	Reports Assessments and Conferences, Failures or Missed Opportunities

Table 6.2: Subclasses for the three datasets.

(consisting of both original and artificially generated samples).

Domain Data In addition to the datasets with a limited amount of labels, we also leverage domain-specific corpora (in the form of the original training sets for each

```
                                levelName passages
1  20 NewsGroups                                NA
2   ¦--computers                              2892
3   ¦   ¦--comp.os.ms-windows.misc             577
4   ¦   ¦--comp.graphics                       579
5   ¦   ¦--comp.sys.ibm.pc.hardware            586
6   ¦   ¦--comp.windows.x                      578
7   ¦   °--comp.sys.mac.hardware              572
8   ¦--rec                                    2378
9   ¦   ¦--rec.motorcycles                     594
10  ¦   ¦--rec.autos                           590
11  ¦   ¦--rec.sport.baseball                  596
12  ¦   °--rec.sport.hockey                    598
13  ¦--talk                                   1569
14  ¦   ¦--talk.politics.guns                  544
15  ¦   ¦--talk.politics.misc                  465
16  ¦   °--talk.politics.mideast               560
17  ¦--science                                2368
18  ¦   ¦--sci.med                             591
19  ¦   ¦--sci.space                           592
20  ¦   ¦--sci.crypt                           595
21  ¦   °--sci.electronics                     590
22  ¦--religion                               1453
23  ¦   ¦--talk.religion.misc                  376
24  ¦   ¦--alt.atheism                         478
25  ¦   °--soc.religion.christian              599
26  °--misc.forsale                           571
```

Figure 6.2: 20 Newsgroups class hierarchy.

```
                    levelName passages
1 Toxic Comments                       NA
2  ¦--non-toxic                    143346
3  °--toxic                         16225
4       ¦--moderate toxic           15294
5       ¦--severe toxic              1595
6       ¦--obscene                   8449
7       ¦--threat                     478
8       ¦--insult                    7877
9       °--identity hate             1405
```

Figure 6.3: Toxic comments class hierarchy.

```
                        levelName sentences passages
1  Safeguarding Reports Themes             NA       NA
2   ¦--Contact with Agencies             1616      485
3   ¦--Indicative Behaviour              1354      506
4   ¦--Indicative Circumstances           531      201
5   ¦--Mental Health Issues               392      134
6   ¦   ¦--Children                        10        4
7   ¦   ¦--Victim                         114       36
8   ¦   ¦--Perp                           136       43
9   ¦   °--Suicidal Ideation               90       30
10  °--Reflections                        983      341
```

Figure 6.4: Safeguarding Reports class hierarchy for Mental Health Issues.

dataset, without making use of the labels) with two purposes: (1) analysing the effect on GPT-2 fine-tuned on more data for generating new instances, and (2) recreating a usual scenario in practice, which is having a relatively large unlabelled corpus (e.g., many comments in the toxicity dataset or a large number of newsgroups) but a small

number of annotations.

6.3 Experimental Settings

6.3.1 Text Generation

As mentioned in Section 6.1, we use GPT-2 standard model for generating additional training instances. We fine-tuned the GPT-2 model using the GPT-2 Hugging Face default transformers implementation for fine-tuning [Wolf et al., 2019]. In addition to the pre-trained general-domain model, we fine-tune GPT-2 in each training set as well as per label using 4 epochs and default settings. For generating additional training sequences we used the sampling decoding method presented by Holtzman et al. [2019]. In early experiments with GPT-2, we used greedy search and beam search, however generated sequences lacked diversity. The artificial sequences have the same length as the seed sequences used to generate them. Further, in cases where the seed length exceeds the maximum length allowed for GPT-2 model (1024), we use the first sentence from the seed for generating a sample. Earlier experiments showed that this approach is more beneficial for producing samples semantically similar to their seeds rather than splitting longer seeds into sentences and generating artificial sequences per sentence, which later are combined into paragraphs. Finally, in order to provide more robustness to the generation process, we performed three iterations of generating additional samples and Section 6.5 present the average results from these three iterations.

6.3.2 Classification

As mentioned earlier, we use fastText as our classifier where we use 'softmax' function, 2 grams, and domain-trained word embeddings. In order to learn domain-specific word embedding models we used the corresponding training sets for each dataset by using fastText's skip-gram model, similarly to previous chapters.

Evaluation metrics. As in previous chapters, we report results based on the standard micro-averaged and macro-averaged F1.

6.3.3 Data Augmentation Baselines

For our baselines, we chose methods representative of the other two widely used data augmentation approaches, i.e., word-replacement and sentence-replacement based approach, described further in Section 2.3.8. Specifically, we employ synonym, word embedding and language model based strategies for word replacement, and back-translation for sentence replacement. For implementing the methods, we rely on *TextAttack* [Morris et al., 2020] for the synonym and word embedding approaches, and *nlpaug* [Ma, 2019] for the language model and back-translation. We follow the default configurations for both libraries, where WordNet is used as a thesaurus for synonym replacement, BERT (*bert-uncased-large*) for the language model, and Transformer NMT models [Vaswani et al., 2017] trained over WMT19 English/Germany corpus for back-translation.

6.4 Case Study with Human Experts

We conducted the case study and consequently the expert-guided seed selection strategy (presented in Section 6.1.1) only for the safeguarding domain where the class framework is created by subject-matter experts. Similar to previous chapters, we performed analysis on sentence-level and passage-level in order to evaluate performance of text generation methods for generating short and long sequences.

For the purposes of the experiments, we randomly selected two samples from the original data, one consisting of sentences ('sentence sample') and another one consisting of passages ('passage sample'), where passages in the safeguarding reports are a list of a few sentences which could be viewed as short paragraphs). Each sample con-

tained 20 instances per label or 100 instances in total. The 'sentence sample' and the 'passage sample' were distributed among two experts (an example of a file with the samples is given in Figure 6.5). Participants were asked for each sentence/passage to choose whether it is a *good representative* of a class or *bad representative* of the class, or to leave space blank if they do not know. The experts followed standard procedures in thematic analysis for completing the task, similar to those used for annotating the safeguarding reports, described in Section 3.1.2. Specifically, they made their choices through discussion. We use only a sample of the original data and involve a small number of experts in order to evaluate whether expert-guided seed selection strategy work in a real case scenario where domain experts are sparse and the selection process is time- and cost- consuming for larger datasets.

theme	text id	text	choice
Reflections	1	In line with the All Wales Child Protection Procedures, a section 47 enquiry was initiated on the 8th December 2010. The section 47 enquiry was completed and closed the same day. It recommended that Chelsey receive some further work .	
Reflections	2	The Specialist Registrar said that he asked the Victim to buy a safe so she can keep his tablets there and then administer them to him, putting some responsibility on her to monitor his medication. His mood stabilizing medication was increased	

Figure 6.5: Example of the file distributed among the experts with non-verbatim examples of the original text.

The results from the experiments, presented in Table 6.3, show that experts have selected more than 10 instances per theme for both samples as 'good representatives'. In order to select 10 and 5 instances from the 'good representatives' we use random selection and max-noun selection where we select the instances with the maximum number of nouns.

Theme	passages		sentences	
	#good representatives	#bad representatives	#good representatives	#bad representatives
Contact with Agencies	12	8	13	7
Indicative Behaviour	12	8	15	5
Indicative Circumstances	11	9	13	7
Mental Health Issues	11	9	14	6
Reflections	11	9	11	9
Total	57	43	66	34

Table 6.3: Results from expert study, where *'#good representatives'* refers to the number of instances selected by the experts as good representatives of the given class while *'#bad representatives'* refers to number of bad representatives of the given class.

6.5 Results and Analysis

The aims of our analyses are (1) to identify the most suitable method for fine-tuning GPT-2 model to ensure generating higher quality training data (see Section 6.5.1), and (2) to understand whether and which seed selection strategies are beneficial for classification performance (see Sections 6.5.2 and 6.5.3).

6.5.1 Can GPT-based Data Augmentation Help Few-shot Text Classification?

Analyses comparing different methods for fine-tuning GPT-2 models for DA for classification showed that GPT-2 model fine-tuned per label lead to higher results for all three datasets, compared to GPT-2 model trained on the entire dataset and pre-trained model (see Tables 6.5 and 6.6). Surprisingly, especially for the more generic datasets such as 20 Newsgroups and Toxic comments, pre-trained GPT-2 model outperforms GPT-2 model fine-tuned on the entire dataset. For instance, for the 20 Newsgroups pre-trained model for '5+5' has micro-F1 = 0.539 while for the same setting fine-tuned model on the domain has micro-F1 = 0.526. This is the case, because a fine-tuned

model without using label-preservation techniques leads to label-distortions which add noise in the generated dataset [2].

The results for the safeguarding reports (see Table 6.6) show a similar trend where the pre-trained model outperforms the model fine-tuned on the entire dataset for most of the settings, except when setting '5base+5add' is used. This is not the case for analysis performed on sentence-level where the model fine-tuned on the entire dataset performs very similarly to the model fine-tuned per label. This further confirms that having a label-preserving techniques in place is highly important for the quality of generated data, especially when longer sequences are used. Further, these results show that fine-tuning the GPT-2 model on smaller but labelled data is more beneficial for classification than a fine-tuned model on a larger unlabelled corpus.

Statistical Significance Test * We used t-test [Student, 1908] to measure whether TG-based DA give a significant improvement over the non-augmented classifiers. T-test is used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is often used to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another. Here, we perform comparison between best performing techniques, which are all based on GPT-2 models fine-tuned per label and the base classifier ('None' in Tables 6.5 and 6.6). We use as a threshold α = 0.05. Results, presented in Table 6.4, showed that $p_{value} < \alpha$ for every dataset. This confirms that fine-tuning a GPT-2 model with a small number of labelled instances versus fine-tuned on larger but unlabelled corpus leads to consistent classification improvement for all datasets.

[2] In Appendix A, Section 7.6 we include automatically-generated samples, comparing the different GPT-2 fine-tuning strategies.

Dataset	p_{micro}	p_{macro}	α
20 Newsgroups	0.01	0.02	0.05
Toxic comments	0.03	0.03	0.05
Safeguarding Reports (passages)	0.0001	0.0001	0.05
Safeguarding Reports (sentences)	0.006	0.016	0.05

Table 6.4: T-test results - comparison between classifier with no augmented data and best performing classifiers with augmented training dataset.

6.5.2 Seed Selection Strategies for Specialised Domains

Figures 6.6 and 6.7 show that the use of any seed selection strategies lead to higher classification results versus random seed selection for both passages and sentences. For instance, for settings '5base + 5add' on passage-level, random seed selection strategy has micro-F1 = 0.295 while the noun-based seed selection has F1-micro = 0.358. However, for passages, noun-based and subclass-guided seed selection, perform worse than the embeddings-based baseline for '5base + 5add' settings, while the use of any of the seed selection strategies on sentence-level lead to higher results than baseline approaches. This suggests that text generation techniques perform better for generating shorter sequences than passages (see Table 6.6)[3].

The case study in the safeguarding reports (see Section 6.4) revealed that seed selection strategy guided by experts outperform all other seed strategies and baselines for both sentences and passages (see Figures 6.6 and 6.7). This highlights the potential benefits for incorporating expert knowledge into guiding large pre-trained language models in specialised domains.

6.5.3 Seed Selection Strategies for Generic Domains

We name 'generic domains' the domains which do not require subject-matter experts knowledge for annotation. Such domains are the '20 newsgroups' and 'toxic com-

[3]In Appendix A, Section 7.7 we include automatically-generated samples for all datasets, showing performance of different seed selection strategies.

	DA type	Tuning type	DA method	**Micro-F1**				**Macro-F1**			
				5base		10base		5base		10base	
				+5add	+10add	+10add	+20add	+5add	+10add	+10add	+20add
20 Newsgroups	None	-	-	.509		.578		.481		.567	
	TG (GPT2)	gen	random	.539	.536	.572	.555	.519	.519	.564	.548
		dom	random	.526	.502	.548	.539	.511	.485	.534	.526
		label*	random	**.609***	**.602***	**.627**	**.637***	**.591***	**.587***	.615	.627
			nouns	.569	.549	.599	.576	.552	.533	.583	.562
			subclass	.563	.585	.624	.632	.549	.571	**.620***	**.628***
	WR	-	BERT	.519	.516	.567	.571	.511	.505	.554	.556
		-	embeddings	.556	.540	.556	.552	.534	.516	.544	.539
		-	synonyms	.517	.508	.554	.549	.502	.493	.542	.537
	SR	-	translation	.529	.525	.559	.563	.515	.509	.549	.552
Original data (upperbound)				.601	.641	.648	.654	.589	.624	.633	.639
Toxic comments	None	-	-	.423		.442		.423		.442	
	TG (GPT2)	gen	random	.447	.424	.405	.423	.447	.424	.405	.423
		dom	random	.401	.417	.369	.343	.401	.417	.369	.343
		label*	random	**.453***	**.452***	.453	.442	**.453***	**.452***	.453	.442
			nouns	.417	.399	**.502***	**.461***	.417	.399	**.502***	**.461***
			subclass	.427	.440	.419	.421	.427	.440	.419	.421
	WR	-	BERT	.447	.443	.426	.422	.447	.443	.426	.422
		-	embeddings	.441	.441	.432	.432	.441	.441	.432	.432
		-	synonyms	.423	.411	.433	.429	.423	.411	.433	.429
	SR	-	translation	.446	-	.436	-	.446	-	.436	-
Original data (upperbound)				.442	.435	.448	.463	.442	.435	.448	.463

Table 6.5: fasText classification results based on Micro-F1 and Macro-F1. Text generation is based on GPT-2, where 'gen' refers to the pre-trained general-domain model, 'dom' refers to the same model fine-tuned on domain data, and 'label', fine-tuned per label. Data is split using 5 or 10 'base' instances per label plus additional 5, 10, or 20 'add' instances. The baselines we compare our approaches to are: the word-based replacement (WR) and sentence-based replacement (SR) strategies (*DA methods based on GPT-2 model fine-tuned per label lead to notable improvements over non-augmented classification (*'None'*) based on t-test results where p_{value} < 0.05).

ments'. Results from comparing seed selection strategies for these datasets showed that random selection especially for smaller amount of seeds is sufficient (see Figures 6.8 and 6.9) for improving classification performance over baselines. However, when larger number of seeds are used ('10 base') and more data is generated from

	DA type	Tuning type	DA method	Micro-F1				Macro-F1			
				5base		10base		5base		10base	
				+5add	+10add	+10add	+20add	+5add	+10add	+10add	+20add
passages	None	-	-	.326		.326		.299		.300	
	TG (GPT2)	gen	random	.298	.305	.382	.358	.254	.264	.335	.330
		dom	random	.333	.288	.323	.309	.276	.246	.287	.267
		label	random	.316	.302	.347	.326	.278	.266	.309	.287
			nouns	.375	.337	.375	.379	.329	.281	.338	.351
			subclass	.379	.330	.368	.368	.321	.286	.335	.345
			expert-random	**.404***	.386	.393	**.407***	**.358***	.349	.342	.352
			expert-nouns	.389	**.435***	**.410***	**.407***	.335	**.382***	**.351***	**.366***
	WR	-	BERT	.287	.294	.326	.336	.282	.278	.294	.297
		-	embeddings	.389	.382	.305	.319	.343	.341	.283	.287
		-	synonyms	.277	.267	.312	.315	.256	.245	.285	.292
	SR	-	translation	.333	.336	.298	.312	.294	.301	.273	.286
Original data (upperbound)				.336	.337	.358	.368	.301	.304	.307	.320
sentences	None	-	-	.242		.316		.193		.282	
	TG (GPT2)	gen	random	.294	.326	.291	.298	.212	.235	.252	.251
		dom	random	.298	.326	.291	.302	.214	.236	.252	.250
		label	random	.295	.326	.291	.302	.213	.235	.251	.252
			nouns	**.358**	.368	.361	.389	.285	.302	.327	.358
			subclass	.330	.351	**.372***	.329	.281	.301	.338	.290
			expert-random	.337	**.375***	.361	**.414***	**.298***	**.336***	**.340***	**.379***
			expert-nouns	.291	.298	.354	.375	.274	.276	.332	.351
	WR	-	BERT	.249	.284	.319	.315	.245	.274	.278	.274
		-	embeddings	.242	.280	.316	.319	.226	.259	.276	.283
		-	synonyms	.256	.266	.319	.326	.241	.256	.281	.288
	SR	-	translation	.287	.294	.336	.329	.257	.263	.296	.291
Original data (upperbound)				.368	.452	.432	.453	.332	.386	.386	.389

Table 6.6: fasText classification results based on Micro-F1 and Macro-F1. Text generation is based on GPT-2, where 'gen' refers to the pre-trained general-domain model, 'dom' refers to the same model fine-tuned on domain data, and 'label', fine-tuned per label. Data is split using 5 or 10 'base' instances per label plus additional 5, 10, or 20 'add' instances. The baselines we compare our approaches to are: the word-based replacement (WR) and sentence-based replacement (SR) strategies (*DA methods based on GPT-2 model fine-tuned per label lead to notable improvements over non-augmented classification (*'None'*) based on t-test results where p_{value} < 0.05).

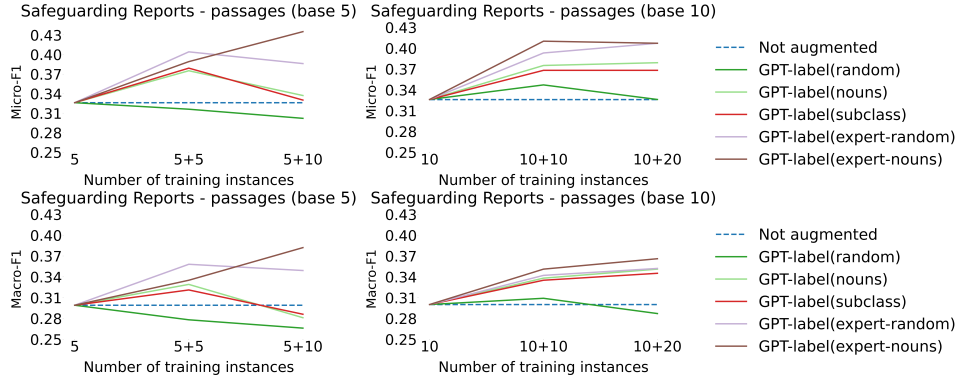

Figure 6.6: Micro-F1 and Macro-F1 results with 5 and 10 'base' instances per label for the Safeguarding reports dataset on passage level.

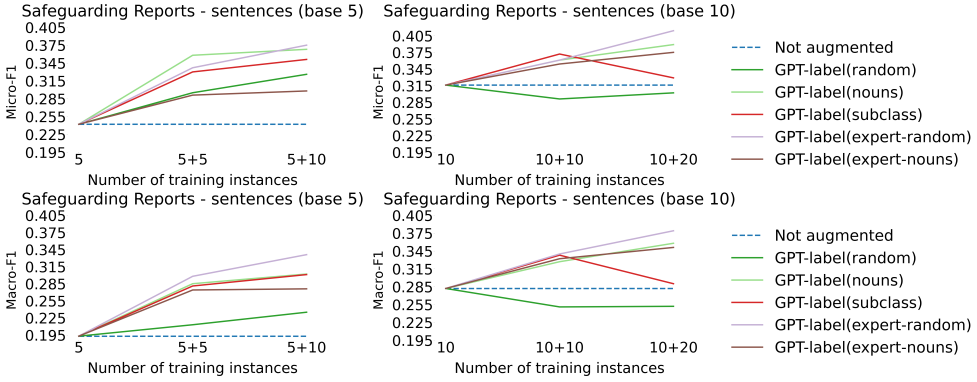

Figure 6.7: Micro-F1 and Macro-F1 results with 5 and 10 'base' instances per label for the Safeguarding reports dataset on sentence level.

these seeds using a selection strategy help improve classification performance. For instance, results for the toxic comments (see Figure 6.9) showed that for 10 base instances the max nouns-based strategy outperforms random selection with around 0.5 improvement in F1 measure with 5 additional instances and 0.2 improvement in F1-measure with 10 additional instances. This suggests that seed selection strategies for more generic domains might be beneficial when larger number of additional training samples are generated. However, further analysis need to be performed to verify this statement.

In contrast to the more specialised domain, i.e., safeguarding reports, text-generation techniques applied to generic domains, combined with random seed selection, outperform all baselines even for small settings. The reason for this is, that the newsgroup and Wikipedia comments are datasets similar to the domains used to train GPT-2 model and therefore the model is more fitted for the 20 Newsgroups and Toxic comments datasets.

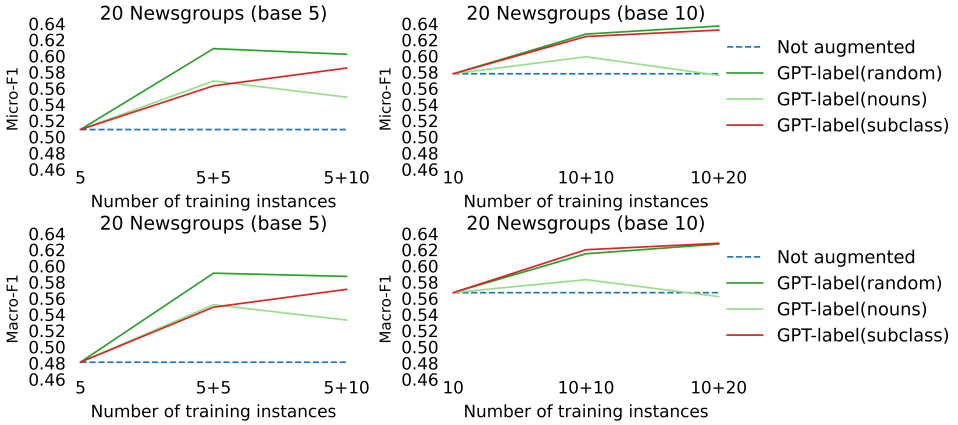

Figure 6.8: Micro-F1 and Macro- F1 results with 5 and 10 'base' instances per label for the 20 Newsgroup dataset.

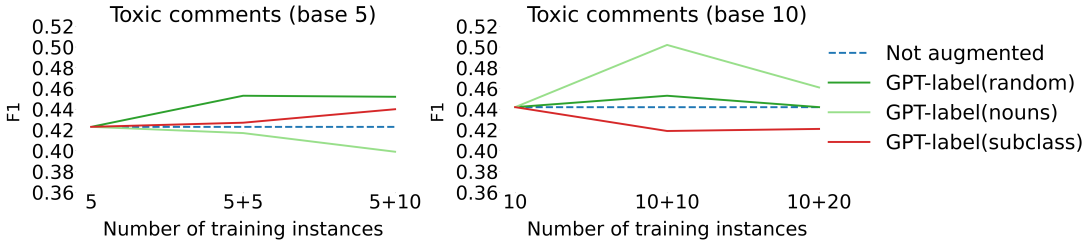

Figure 6.9: F1 results with 5 and 10 'base' instances per label for the Toxic comments dataset.

6.6 Discussion

6.6.1 Data Augmentation Methodology

The methodology consisted of three main steps - Seed Selection, Text Generation, and Classification. We introduced four seed selection strategies in an attempt to improve performance of TG-based DA methods for few-shot text classification focusing mainly on specialised domains. We compared three GPT-2 models, pre-trained, fine-tuned on the entire dataset, and fine-tuned per label in order to identify how different fine-tuning approaches for GPT-2 model affect the quality of generated data.

The seed selection strategies we evaluated were chosen because they leverage characteristics associated with specialised domains. The strategies are: 1) selecting seeds with maximum number of nouns — specialised texts usually have high occurrence of terms and nouns which are indicative for the classes; 2) selecting a roughly balanced number of seeds from each subclass belonging to a given label — ensure instances from underrepresented subclasses participating in the training set; 3) asking an expert to select class representative seeds — applicable to the safeguarding domain where themes have been created by subject-matter experts and thus identifying theme representatives might require implicit knowledge.

We evaluated methods for three datasets from different domains - newsgroups, Wikipedia comments, and the safeguarding reports across four different few shot settings (5base+5add, 5base+10add, 10base+10add, 10base+20add). To ensure robustness in evaluation methods we performed the text generation step of the methodology three times and classification results were averaged for these iterations. Further, we performed a t-test to measure significance of improvement in the classification performance when augmented data generated with GPT-2 is used versus using no augmented data.

6.6.2 Findings

The main findings from this chapter are listed below.

- The use of label-preservation methods when fine-tuning pre-trained text generation models for the classification tasks is important for improving quality of generated data and subsequently classifiers performance — Simply fine-tuning GPT-2 model on the domain data leads to more distortions in the classification performance in comparison to using pre-trained model. This is especially valid for the more generic datasets such as 20 Newsgroups and the toxic comments where the pre-trained model performs significantly better than the model fine-tuned on the entire dataset. In general, the results clearly suggest that fine-tuning the GPT-2 model on smaller but labelled data is equally or more beneficial for classification than fine-tuned model on a larger unlabelled corpus, especially in settings with longer input sequences. This findings also further extend on the conclusions made in the previous chapter, that fine-tuning pre-trained language models to task-specific data, despite the small amount available, is still more beneficial than fine-tuning language models to larger domain-related unlabelled dataset.

- The effectiveness of seed selection strategies is highly dependent on the domain — For the domains, similar to the datasets used to train GPT-2 (Newsgroups and Wikipedia) showed that that random selection especially for smaller amount of seeds is sufficient for improving classification performance over baselines. However, when larger number of seeds are used and more data is generated from these seeds using a selection strategy help improve classification performance. In contrast, applying seed selection techniques to more specialised domain, such as the safeguarding reports, can be highly beneficial for improving classification.

- Seed selection strategies lead to significant improvements over random seed selection for specialised domains — In contrast to more generic domains, for the

safeguarding reports, both seed selection strategies (noun-guided and subclass-guided selection) lead to consistent improvements over random selection even for small number of seed samples. Additionally, seed selection strategies applied to sentences outperformed all baseline approaches. This shows that text-generation techniques might perform better for generating shorter sequences.

- Incorporating expert knowledge into guiding large pre-trained language models is the most suitable method for improving few-shot classification for specialised domains — The use of expert-guided seed selection strategy outperformed all baselines and seed selection strategies even when analysis were performed on passage-level.

6.6.3 Limitations

The main limitation of this research is the lack of further analysis into the performance of text generation models and seed selection strategies when generating higher number of additional training samples. However, the results from this research do show a clear advantage of text generation models, especially when fine-tuned per label, over DA baseline approaches for text classification. Further, seed selection strategies, especially those involving an expert, proved valuable for improving classification for specialised domains.

6.7 Conclusions

In this chapter, we presented data augmentation methods using text generation techniques and seed selection strategies for improving the quality of generated artificial sequences and subsequently classifier's performance in few-shot settings. Specifically, we use these methods to improve the performance of classification models that were found suitable for scarce collections of training data in the previous chapter.

The final research question, **RQ4** from the initial hypothesis was answered in order to show that seed selection strategies especially those involving experts knowledge do help improve the performance of TG-based DA methods for specialised domains. Further, we show that the performance of TG techniques and seed selection strategies is dependent of the domain and amount of data available for fine-tuning. For instance, for more generic domains with large amounts of unlabelled data, similar to these used to train TG model (newsgroups and Wikipedia), seed selection strategies do not give a significant advantage over simple random selection. However, for specialised domains using seed selection strategies lead to improvements for classification in few-shot settings with 5 and 10 seeds. Further, the use of label preservation techniques is shown to be important for the performance of TG-based DA methods, especially for generating longer sequences (such as passages).

The analysis presented in this chapter can be further extended in order to investigate the right balance of generated artificial samples and seed samples. We want to perform analysis with higher number of generated instances per given seed sample and investigate how seed selection strategies perform in such settings. Further, expert-based seed selection strategies can be extended by experimenting with different methods for extracting samples from the original dataset that are used by experts to select seeds.

Chapter 7

Conclusions and Future Work

In this thesis, we aimed to analyse how the performance of text classification techniques can be improved for specialised domains with small volume of data. For these purposes, we used a corpus of safeguarding reports, as an exemplary case study of a specialised domain. In addition to the safeguarding corpus, we used publicly available benchmark datasets to evaluate applicability of methods on a wider scale.

In the initial chapters, we investigated the effectiveness of traditional and less-resource consuming approaches for improving classification performance for specialised domains based on count-based classifiers and feature enrichment methods using lexical databases. The unsatisfactory performance of public lexical databases for the safeguarding domain showed the need for more domain-adaptive and context-aware classification approaches. This motivated a comparison between three classification approaches and various neural network-based feature extraction and feature-integration methods in order to identify suitability of methods for a small corpus of specialised texts. This research showed the potential of state-of-the-art language models, fine-tuned to the task, to perform complex tasks such as thematic analysis, compared to human annotators. However, these analysis also raised questions regarding suitability of state-of-the-art language models for few-shot classification. Therefore, in the next stage of the work, we focused on identifying efficient approaches for few-shot classification. In particular, we conducted quantitative analysis looking at the importance of labelled and unlabelled dataset for few-shot classification. This work showed that linear classification models coupled with domain-trained word embeddings perform

better than larger more data-consuming state-of-the-art language models, pre-trained and domain-trained for few-shot classification. In the latest chapter, we extended on this work by proposing data augmentation methods using text generation techniques and seed selection strategies for improving the quality of generated artificial sequences and subsequently classifier's performance in few-shot settings. Results showed that seed selection strategies are highly beneficial for specialised domains, especially when incorporating expert knowledge into guiding large pre-trained language models. Additionally, using label-preservation techniques for fine-tuning text generation models, even when only a small amount of data is used, is highly beneficial for the performance of DA approaches, regardless of the domain.

In the rest of this chapter we provide an overview and assessment of the work conducted in this thesis, bringing together the ideas from the initial research and how these have helped in developing methods introduced in later chapters. We also discuss how the research presented in the thesis can be taken further in potential future projects. Finally, an overview of the thesis in terms of its contributions is described.

7.1 Analysis of Research and Results

In this thesis, the research behind establishing classification strategies suitable for specialised domains and creating methods for improving the performance of these strategies has been described. In the following sections, we analyse the research that carried out in the primary chapters of this thesis.

7.1.1 Case Study and Exploratory Work: Traditional Information Extraction

We conducted exploratory profiling of the safeguarding dataset by extracting named entities and performing sentiment analysis using a range of well established IE librar-

ies. This work helped identify challenges in analysing the safeguarding reports and identify main characteristics of the dataset.

The highlight of this exploratory work was a research into the existence of traditional and well established classification methods for specialised domains. The use of publicly available lexical databases and ontologies was identified as a widely used approach for enhancing performance for statistical classifiers for specialised domains, especially the medical domain. The less resource-consuming nature of such an approach motivated a research into the suitability of WordNet-based feature enhancement method coupled with count-based classifier for the safeguarding domain. WordNet was selected as a representative of public lexical database due its wide coverage of English terminology. Specifically, we aggregated concepts from WordNet into BOW feature representation by associating low frequency nouns from the corpus with synonym and hypernym concepts from WordNet. We used the enriched BOW vectors as an input to two widely established baseline algorithms in text classification - Naive Bayes and SVM. The WordNet-based augmentation method did not lead to improvements in classification over non-augmented count-based features. These results motivated an investigation into the existence of knowledge graphs which fit the needs of the safeguarding domain. The outcome of this investigations was that there are no publicly available knowledge graphs and lexical resources, suitable for the safeguarding reports.

The results from this explorative analysis on the applicability of existing lexical resources for the safeguarding domain indicated the need to use more domain-targeted methods for performing classification. This triggered a research into the suitability of more context-aware state-of-the-art text representation and classification methods for specialised domains with limited dataset.

7.1.2 Evaluation of State-of-the-art Classification Methods for the Safeguarding Domain

The literature survey on text classification for low resource settings revealed a lack of extensive comparison between different classification approaches and feature representation methods, especially for the purpose of classifying specialised texts. To gain broader understanding on how existing classification methods perform for specialised datasets with limited training data, we compared three supervised approaches — simple linear classification models, and text classification methods based on pre-trained and corpus-trained word embeddings, and state-of-the-art language models. Specifically, we compared the following classifiers — simple linear Naive Bayes model, fastText, and BERT fine-tuned to the task using sequence classification layer. Further, we performed experiments with multiple models and techniques for feature extraction and feature integration. We evaluated the classification models against the annotations generated by the creators of the thematic framework, who we refer to as *expert annotators*. By creating a classifier that uses the annotations generated by *expert annotators* as a 'ground truth', we aimed to produce unified and comparable results across generations that are not susceptible to variations in annotations created by different human annotators interpreting the coding framework. Further to that, in order to evaluate the benefits of using supervised machine learning approaches for annotating documents over manual annotation, we compared classification models performance against the performance of *expert validators*, i.e., independent social scientists who did not participate in the creation of the thematic annotation framework. The comparison was performed on sentence- and passage-level where automated models outperformed expert validators for annotating sentences and performed equally well to the experts for passages. This shows that fine-tuned state-of-the-art language models can perform equally well or even better than expert annotators for labelling specialised documents such as the safeguarding reports. This analysis also suggested that humans need more context — i.e., to see the sentences embedded in paragraphs — to classify sentences

correctly, compared to deep learning models that can generalise better in these cases with limited context thanks to what they learned from the training set. The outcome of this research gives a wider understanding on the usefulness of range of classifiers and feature extraction and integration techniques for performing a complex task, challenging even for humans, such as thematic analysis for specialised documents. Finally, analysis comparing classification approaches for different sized training sets suggested that state-of-the-art language models might be unsuitable for few-shot scenarios.

7.1.3 Suitability of Text Classification Approaches for Few-shot Settings

The conclusions from the previous work motivated further analysis into the effect of training and domain data over the classification performance for different sized datasets, focusing on few-shot settings. Specifically, we performed quantitative analysis comparing linear classification model fastText, coupled with generic and corpus-specific word embeddings, and the pre-trained language model BERT, trained on generic data and domain-specific data. We also included a simple frequency-based classifier for a baseline. The analysis were performed for five domains and six classification tasks, making this research the most extensive comparison between different classification models for few-shot scenarios to date, to the best of our knowledge. Further, we performed experiments by randomly sampling different sized subsets from the original labeled data as well as performing a few-shot experiment where we compared classifier's performance on different sizes of balanced subsets of the training data. The trends into the performance of the classification models show a clear advantage of fastText classifier coupled with domain-trained embeddings over the pre-trained and even domain-trained language model BERT for datasets with less than 2000 training instances. This shows that a simple linear classifier coupled with domain-trained embeddings is more effective for limited sized datasets than a larger more data-consuming language model, pre-trained or trained on domain-related corpus. However, as the

training data size increases, domain-trained and pre-trained BERT models outperform fastText classifier.

This research extends on previous work presented by Gururangan et al. [2020] where authors perform analysis on the effect of pre-training language models (RoBERTa) on the domain and task for various scenarios. However, the author's research is not considering extensive few-shot scenarios involving balanced subsets. Further, it focuses only on evaluating performance of RoBERTa language model rather comparing multiple classification algorithms.

7.1.4 Text Generation-based Data Augmentation Techniques for Few-shot Text Classification

After establishing classification approaches that are suitable for few-shot text classification, we presented a data augmentation methodology using text generation techniques and seed selection strategies for improving performance of these classification approaches. Specifically, we propose a simple data augmentation technique for classification task, based on using seed selection strategies for improving the quality of generated artificial sequences and subsequently classifier's performance in few-shot settings. We proposed four seed selection strategies for selecting class representative samples from the original data used to generate higher quality artificial instances. These are: random selection, subclass-guided selection, max nouns-guided selection, and expert-guided selection. For generating additional samples, we used the GPT-2 model, known to give state-of-the-art performance for various text generation tasks. Additionally, we analysed how different approaches of fine-tuning GPT-2 affect the quality of generated data and consequently the classification performance. Specifically, we compare pre-trained model, fine-tuned model on the entire dataset and fine-tuned model per label. Finally, for a classifier we used fastText coupled with domain-specific embeddings, because this was the approach found most suitable for few-shot classification earlier in the thesis. We performed extensive evaluation of strategies for three

datasets from different domains – newsgroups, Wikipedia comments, and the safeguarding reports across four different few shot settings (5base+5add, 5base+10add, 10base+10add, 10base+20add). Further, we ensured robustness in evaluation methods by performing the text generation step of the methodology three times and averaging classification results for these iterations. Further, we performed a t-test to measure significance of improvement in the classification performance when GPT-2 generated data is used versus using no augmented data. Finally, we used four baselines, three based on word-replacement strategies and one based on back-translation which allows extensive comparison of the proposed approach to other widely used types of data augmentation methods.

In general, analysis showed that data augmentation techniques incorporating GPT-2 model, fine-tuned per label leads to consistent classification improvements across all datasets, compared to the same GPT-2 model fine-tuned on the entire dataset. This highlights the importance of label preservation techniques in the performance of TG-based DA methods, especially for generating longer sequences (such as passages or full documents). The incorporation of seed selection strategies in data augmentation methods showed to be highly beneficial for specialised domains, especially when expert is involved in the seed selection process. However, for the domains similar to the datasets used to train GPT-2 (Newsgroups and Wikipedia), seed selection strategies do not lead to consistent improvements over a simple random selection for small number of seeds.

Previous work by [Yang et al., 2020] proposed an approach based on the use of influence functions and heuristics for selecting the most diverse and informative artificial samples from an already-generated artificial dataset in order to improve quality of artificial training data for classification. The authors performed experiments for common sense reasoning dataset and use GPT-2 model for generating data. The data augmentation methodology based on seed selection strategies for improving data augmentation for text classification is complementing the approach of [Yang et al., 2020] in a number

of ways. Firstly, the research in Yang et al. [2020] is focused on a single domain for common sense reasoning while we evaluate our method for three domains. Second, we focus on the previous step of selecting the most informative samples (or seeds) from the original data. We show that a careful selection of class representative samples from the original data in the first place can already lead to important improvements and has an important efficiency advantage as it prevents an unnecessary waste of resources and time of generating unused generated documents, especially considering how resource expensive generative language models are.

7.2 Contributions

Throughout the earlier chapters of this thesis, work has been conducted towards answering the hypothetical questions asked in Section 1.2 in the Introduction. The research has highlighted the potential of smaller but domain-adapted embedding models for few-shot classification over the use of state-of-the-art language models pretrained on generic or domain relevant data. Further, the research shows that data augmentation methodology based on the text generation model, fine-tuned per label, and seed selection strategies incorporating expert knowledge, do lead to consisted classification improvements in few-shot settings for specialised domains when compared to random seed selection and a range of baseline data augmentation methods. In particular, the questions are now answered more formally.

- **RQ 1: Can publicly available lexical resources be used to support supervised learning for specialised domains?** — Publicly available lexical resources such as WordNet cannot be used to augment statistical classification models coupled with BOW feature representation for specialised domains such as the safeguarding domain.
- **RQ 2: Which classification approaches help preserve subject-matter expert knowledge for annotating specialised unstructured texts, compared to**

human annotators? — State-of-the-art language model such as BERT, fine-tuned to the classification task using sequential classifier outperform or perform equally well as expert annotators, who have not participated in the creation of the thematic framework, for annotating the safeguarding reports. This shows that state-of-the-art deep learning models have the potential to perform complex tasks such as thematic analysis for terminology-rich documents.

- **RQ 3: What are the most efficient approaches for few-shot classification in general and for specialised domains?** — Simple linear classifier such as fastText coupled with domain-trained word embeddings outperformed pre-trained and domain-trained language model such as BERT for classification with small volume of training data for both few-shot scenarios with a balanced set and keeping the original distributions, regardless of the domain used.

- **RQ 4: Can text classification performance be improved through the use of data augmentation techniques based on text generation and seed selection strategies in few-shot settings in general and for specialised domains?** — In general, data augmentation techniques based on text generation model, fine-tuned per label achieve a consistent improvements in few-shot text classification, compared to baseline approaches and the same text generation model fine-tuned to the entire dataset. Guiding the text generation process using seed selection strategies for data augmentation proved highly beneficial for specialised domains, especially when experts participate in the seed selection process. For domains which are more similar to the datasets used to train text generation model (Newsgroups and Wikipedia), seed selection strategies do not lead to consistent improvements over a simple random selection for small number of seeds. However, for larger numbers of seeds, strategies do help classification performance.

7.3 Future Work

7.3.1 Extend on Quantitative Analysis

The quantitative analysis presented in Chapter 5 can be further extended by performing analysis for larger language models such as RoBERTa, more datasets and by further evaluating the role of unlabeled data for text classification by using meta-embeddings. A simple method for creating meta-embeddings can be concatenating BERT and domain-trained fastText embeddings. Also, analysis can be extended by evaluating performance of language models when they are continuously pre-trained on the domain and task.

7.3.2 Extend on TG-DA Methodologies for Text Classification

The TG-based DA methodology for text classification, presented in Chapter 6 need to be extended considering higher number of generated additional sequences. We plan to investigate the optimal number of generated instances using GPT 2-based generation. Further, we want to extend analysis for more classification tasks and datasets. Moreover, given the positive results from the expert guided generation, we plan on exploring more methods involving human expertise into the seed selection process.

7.3.3 Adaptive Hierarchical Classification

Throughout this thesis we focused on the problem of identifying classification methods and improving them using data augmentation techniques for specialised domains with limited training data. As a case study we used the task of automating the thematic analysis for a small collection of safeguarding reports focusing on predicting the five overall themes of the thematic framework. A logical continuation of the project will be to develop a hierarchical classification for predicting all themes within the thematic

framework. The framework consists in total of 100 themes and sub-themes where some leave themes are associated with less than 3 samples. This makes the task of building a hierarchical classifier for predicting all themes a challenging task with important applications for many domains with hierarchical organisation of labels and limited data.

We are planning on using a top-to-bottom approach for performing hierarchical classification where we first perform classification for the overall classes and then use weight-propagation techniques for predicting sub-themes. As TG-based DA techniques have been proved in the thesis to help text classification for specialised domains, we plan on incorporating these techniques within the hierarchical classification process.

A problem with many real-world scenarios is that classification frameworks are susceptible to frequent changes Zhang et al. [2019], Ye et al. [2020], Chalkidis et al. [2020a] such as insertion, deletion or change of some of the classes. For instance, for the safeguarding domain, it is common when new documents are introduced to the collection, for new problems to be identified which causes change of themes, insertion of new themes, or merges of themes. Therefore, an interesting continuation of creating a hierarchical classifier for predicting themes will be to also support its adaption to changes for zero-shot learning.

The creation of a hierarchical classification approach adaptable to zero-shot settings can also facilitate the creation of more accurate and efficient automated tools for the WSR for annotating new safeguarding reports with themes.

7.3.4 Develop Semantic Search Tool for the Safeguarding Domain

As mentioned in the beginning of the thesis, one of the most challenging aspects of analysing the safeguarding reports is the diverse and highly specialised terminology used within the corpus. Further, writing as well as analysing the reports involve the

participation of practitioners and researchers with different expertise and background which also reflects on the language they use for describing the same concept, i.e., 'hoodwinking' and 'lying', as well as 'coercive control' and 'domestic violence'. Thus, in order to support conducting cross-report and cross-collection analysis, there is need to extend the Wales Safeguarding Repository by providing semantic search tools which support the identification of similar problems despite the diverse terminology used to describe them. This can be achieved by further enhancing the hierarchical classification of themes by creating domain-specific lexical resources.

7.4 Final Remarks

The work in this thesis was carried out with the aim of researching methods that facilitate text classification for specialised domains with limited amount of data and terminology, not widely assembled in existing lexical resources and pre-trained neural models. Initially, we performed a study on more traditional but less resource-consuming approaches for enhancing classification performance based on feature enrichment using lexical resources method coupled with statistical ML algorithms. The findings from this research showed the need of more context-aware approaches for performing classification in specialised domains. This motivated the conduct of thorough analyses of state-of-the-art classification approaches with a focus on NN-based models and their suitability for small and specialised corpora. We particularly focused on investigating the affect of labelled and unlabelled data over state-of-the-art classification models in few-shot scenarios. This comparison showed that simpler and less data consuming linear models coupled with domain-trained embeddings are more suitable for small corpus than larger state-of-the-art language models, pre-trained and domain-trained. These analyses helped identify classification strategies applicable to small datasets.

The research processes culminated in the development of a data augmentation methodology that help improve performance of classification strategies in few-shot settings.

The approach consisted of using text generation models and seed selection strategies that facilitate the generation of higher quality additional training data and in turn lead to improvements in classification. Extensive analysis into two fine-tuning approaches of the text generation model and four seed selection strategies, showed that the use of label-preservation techniques for fine-tuning of generation models (even when only a small amount of samples is used) and the incorporation of expert knowledge into the seed selection process is a suitable method for improving few-shot classification for specialised texts, compared to four baselines and other seed selection strategies.

Throughout the thesis we used the safeguarding reports as an exemplary case study of a specialised corpus with limited data. In addition, we used a range of benchmark datasets to allow a robust comparison and evaluation of methods. Additionally, the research work represented by this thesis is being developed further towards applications in WSR project. Specifically,classification models developed throughout this thesis will be integrated into the WSR interface. Further, research will be extended to support semantic search tools as part of WSR that can help practitioners and researchers from health and social sciences into faster and more accurate decision-making.

Bibliography

Charu C Aggarwal. Content-based recommender systems. In *Recommender systems*, pages 139–166. Springer, 2016.

Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

Charu C Aggarwal et al. Neural networks and deep learning. *Springer*, 10:978–3, 2018.

Amanuel Alambo, Cori Lohstroh, Erik Madaus, Swati Padhee, Brandy Foster, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Michael Raymer. Topic-Centric Unsupervised Multi-Document Summarization of Scientific and News Articles. *arXiv preprint arXiv:2011.08072*, 2020.

Zuhair Ali. Text classification based on fuzzy radial basis function. *Iraqi Journal for Computers and Informatics*, 45(1):11–14, 2019.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://www.aclweb.org/anthology/W19-1909`.

Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

Berna Altinel and Murat Can Ganiz. Semantic text classification: A survey of past and recent advances. *Information Processing Management*, 54:1129–1153, 2018. doi: https://doi.org/10.1016/j.ipm.2018.08.001.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. Exploring Transformer Text Generation for Medical Dataset Augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.578.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do Not Have Enough Data? Deep Learning to the Rescue! In *Proceedings of AAAI*, pages 7383–7390, 2020.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Ashutosh Baheti, Alan Ritter, and Kevin Small. Fluent Response Generation for Conversational Question Answering. *arXiv preprint arXiv:2005.10464*, 2020.

Katherine Bailey and Sunny Chopra. Few-shot text classification with pre-trained word embeddings and a human in the loop. *arXiv preprint arXiv:1804.02063*, 2018.

Satanjeev Banerjee and Ted Pedersen. An adapted Lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, pages 136–145, Berlin, Heidelberg, 2002. Springer.

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5108–5123, 2020.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States, 2018. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003.

H Russell Bernard and Harvey Russell Bernard. *Social research methods: Qualitative and quantitative approaches*. Sage, 2013.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

Thorsten Brants, Ashok C Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large Language Models in Machine Translation. In *EMNLP-CoNLL*, 2007.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, pages http–openreview, 2014.

Jose Camacho Collados, Yerai Doval, Eugenio Martínez-Cámara, Luis Espinosa-Anke, Francesco Barbieri, and Steven Schockaert. Learning cross-lingual word embeddings from Twitter via distant supervision. *ICWSM*, 2020.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. IMHO Fine-Tuning Improves Claim Detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1054. URL `https://www.aclweb.org/anthology/N19-1054`.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, 2020a.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020b.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. Importance of Semantic Representation: Dataless Classification. In *Aaai*, volume 2, pages 830–835, 2008.

Michael Chau, Jennifer J Xu, and Hsinchun Chen. Extracting meaningful entities from police narrative reports. In *Proceedings of the 2002 annual national conference on Digital government research*, pages 1–5. Digital Government Society of North America, 2002.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, 2017.

Yung-Yao Chen, Yu-Hsiu Lin, Chia-Ching Kung, Ming-Han Chung, I Yen, et al. Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes. *Sensors*, 19(9):2047, 2019.

Hsiao-Yu Chiang, Jose Camacho-Collados, and Zachary Pardos. Understanding the Source of Semantic Regularities in Word Embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 119–131, 2020.

Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. GRAM: graph-based attention model for healthcare representation learning.

In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795, 2017.

D Manning Christopher, Raghavan Prabhakar, Schütze Hinrich, et al. Introduction to information retrieval. *An Introduction To Information Retrieval*, 151(177):5, 2008.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1213. URL https://www.aclweb.org/anthology/N19-1213.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10, 2013.

Francesco Colace, Massimo De Santo, Luca Greco, and Paolo Napoletano. Text classification using a few labeled examples. *Computers in Human Behavior*, 30:689–697, 2014.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1070.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1469–1477, 2015.

Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology*, 9(2):e1002854, 2013.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *NIPS*, 2016.

Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. When low resource NLP meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13773–13774, 2020.

Xuelian Deng, Yuqing Li, Jian Weng, and Jilian Zhang. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3):3797–3816, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Fabio Henrique Kiyoiti dos Santos Tanaka and Claus Aranha. Data Augmentation Using GANs. *Proceedings of Machine Learning Research XXX*, 1:16, 2019.

Xinya Du, Junru Shao, and Claire Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, 2017.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, 1998.

Aleksandra Edwards, Alun Preece, and Helene De Ribaupierre. Knowledge Extraction from a Small Corpus of Unstructured Safeguarding Reports. In *European Semantic Web Conference*, pages 38–42, PortoroÅº, Slovenia, 2019. Springer.

Aleksandra Edwards, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. Go Simple and Pre-Train on Domain-Specific Corpora: On the Role of Training Data for Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, 2020.

Aleksandra Edwards, David Rogers, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. Predicting Themes within Complex Unstructured Texts: A Case Study on Safeguarding Report. In *2nd International Workshop Deep Learning meets Ontologies and Natural Language Processing, European Semantic Web Conference (ESWC 2021)*, 2021a.

Aleksandra Edwards, Asahi Ushio, Jose Camacho-Collados, Hélène De Ribaupierre, and Alun Preece. Guiding Generative Pre-trained Language Models for Data Augmentation in Few-Shot Text Classification. In *Submitted to the 2021 Conference on Emirical Methods in Natural Language Processing (EMNLP 2021)*, 2021b.

Kawin Ethayarajh. Unsupervised random walk sentence embeddings: A strong but simple baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100, Melbourne, Australia, 2018.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*, 2017.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, 2013.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. Retrofitting Word Vectors to Semantic Lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1184. URL `https://www.aclweb.org/anthology/N15-1184`.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

Jerome A Feldman and Dana H Ballard. Connectionist models and their properties. *Cognitive science*, 6(3):205–254, 1982.

Diego Fernandes de Araújo, Carlos Eduardo Santos Pires, and Dimas Cassimiro Nascimento. Leveraging active learning to reduce human effort in the generation of ground-truth for entity resolution. *Computational Intelligence*, 36(2):743–772, 2020.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6407–6414, 2019.

Raphaël Gazzotti, Catherine Faron-Zucker, Fabien Gandon, Virginie Lacroix-Hugues, and David Darmon. Injecting Domain Knowledge in Electronic Medical Records to Improve Hospitalization Prediction. In *European Semantic Web Conference*, pages 116–130. Springer, 2019.

Raphaël Gazzotti, Catherine Faron-Zucker, Fabien Gandon, Virginie Lacroix-Hugues, and David Darmon. Injection of automatically selected DBpedia subjects in electronic medical records to boost hospitalization prediction. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2013–2020, 2020.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. Induction Networks for Few-Shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Interna-*

tional Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3895–3904, 2019.

Anna Lisa Gentile, Daniel Gruhl, Petar Ristoski, and Steve Welch. Explore and exploit. Dictionary expansion with human-in-the-loop. In *European Semantic Web Conference*, pages 131–145. Springer, 2019.

Spyros Gidaris and Nikos Komodakis. Generating classification weights with GNN denoising autoencoders for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2019.

Praveen Kumar Badimala Giridhara, Chinmaya Mishra, Reddy Kumar Modam Venkataramana, Syed Saqib Bukhari, and Andreas Dengel. A Study of Various Text Augmentation Techniques for Relation Classification in Free text. *ICPRAM*, 3:5, 2019.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Yoav Goldberg. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309, 2017.

Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. TableGPT: Few-shot Table-to-Text Generation with Table Structure Reconstruction and Content Matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988, 2020.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005.

Aakriti Gupta, Kapil Thadani, and Neil O'Hare. Effective Few-Shot Classification with Transfer Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.92. URL https://www.aclweb.org/anthology/2020.coling-main.92.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

Xiaochuang Han and Jacob Eisenstein. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1433. URL https://www.aclweb.org/anthology/D19-1433.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. Findings of the Third Workshop on Neural Generation and Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, 2019.

Bradford Heap, Michael Bain, Wayne Wobcke, Alfred Krzywicki, and Susanne Schmeidl. Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems. *arXiv preprint arXiv:1709.05778*, 2017.

Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

GE Hinton, JL McClelland, and DE Rumelhart. Distributed representations. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, volume 1, pages 77–109. MIT Press, 1986.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2019.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. Sequence-to-sequence Data Augmentation for Dialogue Language Understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, 2018.

Jeremy Howard and Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, 2018.

Fei Hu, Li Li, Xiaofei Xu, Jingyuan Wang, and Jinjing Zhang. Opinion extraction by distinguishing term dependencies and digging deep text features. *Neural Computing and Applications*, 31(9):5419–5429, 2019.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Mohammed J Islam, QM Jonathan Wu, Majid Ahmadi, and Maher A Sid-Ahmed. Investigating the Performance of Naive-Bayes Classifiers and K-Nearest Neighbor Classifiers. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, pages 1541–1546. IEEE, 2007.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106 (4):620, 1957.

Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, 2014.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs RNN in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.

Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. Automatic Text Summarization of COVID-19 Medical Research Articles using BERT and GPT-2. *arXiv preprint arXiv:2006.01997*, 2020.

Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tassilo Klein and Moin Nabi. Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. *arXiv preprint arXiv:1911.02365*, 2019.

Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1):1–40, 2016.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Sosuke Kobayashi. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

Yushen Kong, Micheal Owusu-Akomeah, Henry Asante Antwi, Xuhua Hu, and Patrick Acheampong. Evaluation of the robusticity of mutual fund performance in ghana using enhanced resilient backpropagation neural network (ERBPNN) and fast adaptive neural network classifier (FANNC). *Financial Innovation*, 5(1):1–12, 2019.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

Saranya Krishnan and Min Chen. Identifying tweets with fake news. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 460–464. IEEE, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, 60:6, 2017.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020.

Ichiro Kuriki, Ryan Lange, Yumiko Muto, Angela M Brown, Kazuho Fukuda, Rumi Tokunaga, Delwin T Lindsey, Keiji Uchikawa, and Satoshi Shioiri. The modern Japanese color lexicon. *Journal of vision*, 17(3):1–1, 2017.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

SM Lakew, M Cettolo, and M Federico. A comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In *27th International Conference on Computational Linguistics (COLING)*, pages 641–652, 2018.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339, Tahoe City, California, 1995.

Eitel JM Lauría and Alan D March. Combining Bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis. *Journal of Data and Information Quality (JDIQ)*, 2(3):1–22, 2011.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5 (Apr):361–397, 2004.

Hongmin Li, D Caragea, X Li, and Cornelia Caragea. Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks. *en. In: New Zealand*, page 13, 2018.

Ximing Li and Bo Yang. A pseudo label based dataless Naive Bayes algorithm for text classification with seed words. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1908–1917, 2018.

P. Liu, X. Wang, C. Xiang, and W. Meng. A Survey of Text Data Augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195, 2020. doi: 10.1109/CCNS50731.2020.00049.

Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. A Survey of Text Data Augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195. IEEE, 2020.

Peter J Liu. Learning to write notes in electronic health records. *arXiv preprint arXiv:1808.02622*, 2018.

Shuai Liu and Xiaojun Huang. A Chinese Question Answering System based on GPT. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pages 533–537. IEEE, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1335. URL `https://www.aclweb.org/anthology/P19-1335`.

Yong Luo, Jian Tang, Jun Yan, Chao Xu, and Zheng Chen. Pre-trained multi-view word embedding using two-side neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

Chen Lyu, Weijie Liu, and Ping Wang. Few-Shot Text Classification with Edge-Labeling Graph Neural Network-Based Prototypical Network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5547–5552, 2020.

Edward Ma. NLP Augmentation. https://github.com/makcedward/nlpaug, 2019.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. Controlled Text Generation for Data Augmentation in Intelligent Artificial Agents. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 90–98, 2019.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

Vukosi Marivate and Tshephisho Sefara. Improving short text classification through global augmentation methods. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 385–399. Springer, 2020.

Every Child Matters. Working together to safeguard children. *London: The Stationery Office*, 2006.

Diana Maynard and Adam Funk. Combining expert knowledge with nlp for specialised applications. In Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors, *Text, Speech, and Dialogue*, pages 3–10, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58323-1.

A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *AAAI 1998*, 1998.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6297–6308, 2017.

Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5, 2001.

Oren Melamud and Chaitanya Shivade. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, 2019.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*, 2018.

Danny Merkx and Stefan L Frank. Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv preprint arXiv:2005.09471*, 2020.

Katie Metzler, David A Kim, Nick Allum, and Angella Denman. *Who is doing computational social science? Trends in big data research*, 2016 (accessed February 3, 2014). URL `https://us.sagepub.com/sites/default/files/CompSocSci.pdf`.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013c.

Erik G Miller, Nicholas E Matsakis, and Paul A Viola. Learning from one example through shared densities on transforms. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 464–471. IEEE, 2000.

George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

Hua Min, Hedyeh Mobahi, Katherine Irvin, Sanja Avramovic, and Janusz Wojtusiak. Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. *Journal of biomedical semantics*, 8(1):39, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A Simple Neural Attentive Meta-Learner. In *International Conference on Learning Representations*, 2018.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, 2020.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324, 2017. doi: 10.1162/tacl_a_00063. URL `https://www.aclweb.org/anthology/Q17-1022`.

Nikolaos Mylonas, Stamatis Karlos, and Grigorios Tsoumakas. Zero-Shot Classification of Biomedical Articles with Emerging MeSH Descriptors. In *11th Hellenic Conference on Artificial Intelligence*, pages 175–184, 2020.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.01973*, 2019.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

Alicia L Nobles, Jeffrey J Glenn, Kamran Kowsari, Bethany A Teachman, and Laura E Barnes. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

Bahadorreza Ofoghi and Karin Verspoor. Textual Emotion Classification: An Interoperability Study on Cross-Genre data sets. In *Australasian Joint Conference on Artificial Intelligence*, pages 262–273. Springer, 2017.

Yannis Papanikolaou and Andrea Pierleoni. Data augmented relation extraction (DARE) with GPT-2. *Neuropharmacology*, 2019.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. Hierarchical Transformers for Long Document Classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://www.aclweb.org/anthology/N18-1202`.

Matthew E Peters, Sebastian Ruder, and Noah A Smith. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, 2019.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. Towards a Seamless Integration of Word Senses into Downstream NLP Applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1857–1869, 2017.

Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard de Melo, Chong Long, and Xiaolong Li. Easyaug: An automatic textual data augmentation platform for classification tasks. In *Companion Proceedings of the Web Conference 2020*, pages 249–252, 2020.

Care Quality Commission. Safeguarding people, 2014. URL `https://www.cqc.org.uk/what-we-do/how-we-do-our-job/safeguarding-people`.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *journal*, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

T. C. Rajapakse. Simple Transformers. `https://github.com/ThilinaRajapakse/simpletransformers`, 2019.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *Proceedings of the International Conference on Learning Representations*, 2017.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 991–1000, 2019.

Amanda Lea Robinson, Alyson Rees, and Roxanna Dehaghani. Making connections: a multi-disciplinary analysis of domestic homicide, mental health homicide and adult practice reviews. *The Journal of Adult Protection*, 2019.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *arXiv preprint arXiv:2002.12327*, 2020.

Fabio Roli, Giorgio Giacinto, and Gianni Vernazza. Comparison and combination of statistical and neural network algorithms for remote-sensing image classification. In *Neurocomputation in remote sensing data analysis*, pages 117–124. Springer, 1997.

Magnus Sahlgren and Alessandro Lenci. The effects of data size and frequency range on distributional semantic models. *arXiv preprint arXiv:1609.08293*, 2016.

Oscar Sainz and German Rigau. Ask2Transformers: Zero-Shot Domain labelling with Pre-trained Language Models. *arXiv preprint arXiv:2101.02661*, 2021.

Alberto G Salguero, Macarena Espinilla, Pablo Delatorre, and Javier Medina. Using ontologies for the online recognition of activities of daily living. *Sensors*, 18(4):1202, 2018.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Yuta Sano, Kohei Yamaguchi, and Tsunenori Mine. Automatic classification of complaint reports about city park. *Information Engineering Express*, 1(4):119–130, 2015.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.

Victor Garcia Satorras and Joan Bruna Estrach. Few-Shot Learning with Graph Neural Networks. In *International Conference on Learning Representations*, 2018.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few-shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *arXiv preprint arXiv:1907.10597*, 2019.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*, 2015.

Rushdi Shams. Semi-supervised classification for natural language processing. *arXiv preprint arXiv:1409.7612*, 2014.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 107–114, 2018.

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327, 2018.

Sonit Singh. Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*, 2018.

Roberta A Sinoara, Jose Camacho-Collados, Rafael G Rossi, Roberto Navigli, and Solange O Rezende. Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163:955–971, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090, 2017.

Richard Socher, Eric H Huang, Jeffrey Pennington, Andrew Y Ng, and Christopher D Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, volume 24, pages 801–809, 2011.

Xingyi Song, Johnny Downs, Sumithra Velupillai, Rachel Holden, Maxim Kikoler, Kalina Bontcheva, Rina Dutta, and Angus Roberts. Using deep neural networks with intra-and inter-sentence context to classify suicidal behaviour. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1303–1310, 2020.

Yangqiu Song and Dan Roth. On dataless hierarchical text classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Irena Spasić, Mark Greenwood, Alun Preece, Nick Francis, and Glyn Elwyn. Flexiterm: a flexible term recognition method. *Journal of biomedical semantics*, 4(1):27, 2013.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL `https://www.aclweb.org/anthology/P19-1355`.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Cheng Niu, Jie Zhou, et al. Diversifying Dialogue Generation with Non-Conversational Text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097, 2020.

Masashi Sugiyama. *Introduction to statistical machine learning*. Morgan Kaufmann, 2015.

Heung-Il Suk. An introduction to neural networks and deep learning. In *Deep Learning for Medical Image Analysis*, pages 3–24. Elsevier, 2017.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune BERT for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, 27:3104–3112, 2014.

Swabha Swayamdipta, Matthew Peters, Brendan Roof, Chris Dyer, and Noah A Smith. Shallow syntax in deep water. *arXiv preprint arXiv:1908.11047*, 2019.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Liling Tan. Pywsd: Python Implementations of Word Sense Disambiguation (WSD) technologies [software]. https://github.com/alvations/pywsd, 2014.

Jie Tang, Mingcai Hong, Duo Liang Zhang, and Juanzi Li. Information extraction: Methodologies and applications. In *Emerging Technologies of Text Mining: Techniques and Applications*, pages 1–33. IGI Global, 2008.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

Yuta Tsuboi. Neural networks leverage corpus-wide information for part-of-speech tagging. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 938–950, 2014.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010.

Rima Türker. Knowledge-Based Dataless Text Categorization. In Pascal Hitzler, Sabrina Kirrane, Olaf Hartig, Victor de Boer, Maria-Esther Vidal, Maria Maleshkova, Stefan Schlobach, Karl Hammar, Nelia Lasierra, Steffen Stadtmüller, Katja Hose, and Ruben Verborgh, editors, *The Semantic Web: ESWC 2019 Satellite Events*, pages 231–241, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32327-1.

Rima Türker, Lei Zhang, Maria Koutraki, and Harald Sack. Knowledge-based short text categorization using entity and category embedding. In *European Semantic Web Conference*, pages 346–362. Springer, 2019.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

Howard Turtle. Text retrieval in the legal world. *Artificial Intelligence and Law*, 3(1): 5–54, 1995.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3637–3645, 2016.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Proceedings of NeurIPS*, 2019a.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *In the Proceedings of ICLR.*, 2019b. In the Proceedings of ICLR.

Chenguang Wang, Mu Li, and Alexander J Smola. Language models with transformers. *arXiv preprint arXiv:1904.09408*, 2019c.

Congcong Wang and David Lillis. Classification for Crisis-Related Tweets Leveraging Word Embeddings and Data Augmentation. In *TREC*, 2019.

Sida I Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, 2012.

William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, 2015.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

Jason Wei and Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6383–6389, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE, 2016.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. Conditional BERT contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer, 2019.

Liqiang Xiao, Lu Wang, Hao He, and Yaohui Jin. Modeling Content Importance for Summarization with Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3606–3611, 2020.

Yijun Xiao and Kyunghyun Cho. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*, 2016.

Binxia Xu, Siyuan Qiu, Jie Zhang, Yafang Wang, Xiaoyu Shen, and Gerard de Melo. Data Augmentation for Multiclass Utterance Classification–A Systematic Study. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5494–5506, 2020.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. G-DAug: Generative Data Augmentation for Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1008–1025, 2020.

Yiming Yang. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90, 1999.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. *arXiv preprint arXiv:1711.03953*, 2017.

Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. Zero-shot Text Classification via Reinforced Self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3014–3024, 2020.

Wenpeng Yin and Hinrich Schütze. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1351–1360, Berlin, Germany, 2016. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In *International Conference on Learning Representations*, 2018a.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Diverse Few-Shot Text Classification with Multiple Metrics. In *NAACL-HLT*, 2018b.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. On Data Augmentation for Extreme Multi-label Classification. *arXiv preprint arXiv:2009.10778*, 2020a.

Jinghe Zhang, Kamran Kowsari, James H Harrison, Jennifer M Lobo, and Laura E Barnes. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. Integrating Semantic Knowledge to Tackle Zero-shot Text Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 649–657, 2015.

Xiaoya Zhang, Lianjie Wang, Jin Xie, and Pengfei Zhu. Human-in-the-loop image segmentation and annotation. *Science China Information Sciences*, 63(11):1–3, 2020b.

Xinwei Zhang and Bin Wu. Short text classification based on feature extension using the n-gram model. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 710–716. IEEE, 2015.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, 2018.

Xiao Zhong and David Enke. Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, 5(1):1–20, 2019.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer, 2017.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Appendix A

7.5 Original datasets: Description and Results

The full description of the original datasets is given in Table 7.1. Results from performing classification using unmodified datasets (using the full training data) are given in Table 7.2.

Dataset	Domain	Task	Type	Class	Subclass	Avg tokens	# Train	# Test
Safeguarding Reports (passages)	Social science	Theme detection	Document	5	34	45	1,261	284
Safeguarding Reports (sentences)	Social science	Theme detection	Document	5	34	18	3,591	284
20 Newsgroups	Newsgroups	Topic categorization	Document	6	20	285	11,231	6,728
Toxic comments	Wikipedia	Toxic prediction	Document	2	5	46	159,571	63,978

Table 7.1: Description of unmodified datasets used in paper experiments.

Dataset	Micro-F1	Macro-F1
20 Newsgroups	0.768	0.759
Toxic comments	0.908	0.908
Safeguarding Reports (passages)	0.463	0.404
Safeguarding Reports (sentences)	0.505	0.477

Table 7.2: FastText classification results for the entire datasets with with no augmentation.

7.6 Comparison between Fine-tuning Techniques for GPT-2

Table 7.3 provide examples of generated sequences using pre-trained, fine-tuned per label, and fine-tuned per the entire dataset GPT-2 model.

Dataset	Label	Original Text	GPT(gen) generation	GPT(dom) generation	GPT(label) generation
20 Newsgroups	Computers	There are 3 alternatives: You can try to write a small C program something like: int my_int_var;my_int_var = system(command);	4. It's a bit confusing and requires very little programming knowledge.5. It uses only your current CPU and some extra memory.	I am now posting the following FAQs to a group of individuals whose answers are from people who are now either atheists, agnostics, etc.	It appears that I am having a problem in trying to use XRemoteAddInput(xmodmap) in windows.
Toxic Comments	non-toxic	think the recent editing by all has been brilliant. If I see anything that could be improved I'll speak up here. I would like to see a Christian Theosophy section also.	Thank you for your help, H, for your work on my blog.You are welcome to submit comments in this section, but please note I am not allowed to comment on other forums because they may interfere with my work.	No need to be an encyclopedic asshole. You don't see me deleting things because of something that was on my watchlist. I just wanted you to know that my name is now on my watchlist, and I will try to improve it.	I hope to see the article finished as soon as possible. But I would like to thank you for your contributions.
Safeguarding reports	Indicative Circumstances	Medical records show that she struggled for many years with depression, for which she was treated with anti-depressants, also receiving referrals for counselling. She also had issues with her physical health.	Her lawyer, Richard Gage, has asked the court to intervene if Mr. McDevitt was not criminally liable in his decision.	She attended at her GP with a range of depressive symptoms and her medication was improved. It is not known how she gained her weight but it is said that her mental health declined and she started following a diet with vitamin C, which was associated with being better at controlling her mood.	In the early 1970s she moved from Los Angeles to New York. During this time she suffered from depression and other problems, but there was little support in New York.She moved into a small apartment and was living with friends, but had to move into a house next to her apartment to support her mental health.

Table 7.3: Examples of generated samples, comparing different GPT-2 models, where Safeguarding Report examples are non-verbatim due to data sensitivity.

7.7 Examples of Generated Sequences using GPT-2 Model Fine-tuned per Label

Tables 7.4, 7.5, 7.6, and 7.7 present examples of generated sequences using GPT-2 model fine-tuned per label and each one of the seed selection strategies, presented in Chapter 6, i.e., random seed selection, maximum nouns-guided seed selection, subclass-guided seed selection, and expert-guided seed selection.

Dataset	Original Seed	GPT-2 Generation	Label
20 Newsgroups	I am in the market for a new bike (been without for a few years). The two main bikes I'm looking at seriously are The Yamaha Virago 535 and the Honda Shadow VLX 583. I am leaning towards the Yamaha for its shaft drive, the Honda is Chain. Insurance in Fla. is more costly than I thought, so I am staying in this power range. Thanks in advance for any opinions and or experiences	I may have.My bike has been driven by an older 70 Buick Cherokee with an engine rated at 120hp. I have the original engine replaced in the dealer. I was told to check all that came with it and to look for leaks. I am not sure what is wrong with the engine since I cant afford a rebuilt engine. Does anybody know any good advice on making good use of the original transmission?Thanks.	recreation
	Yes, that is most certainly true. However, the paragraph reflects a value-less position and infers that what is more important than anything else is to have a positive view of one's self. This of course, is foolish. Should a mass murderer, a pedophile, a 10-year old pyromaniac have a positive view of themselves? Of course not. A person that engages in behaviour that a large number of people condemn, and IF you believe in the concept of society, then your only choice is to expect that person to have a negative view of themselves. There are actually people that STILL believe Love Canal was some kind of environmental disaster. Weird, eh?	Just wanted to remind you that, in the United States, the BATF has a number of different standards that you should know about. They are:1) A uniform, law-abiding federal law enforcement officer, and the BATF has two different standards in that they can legally arrest people for violating federal law(e.g., possession of a loaded firearm or possession of a concealed weapon) and/or have them serve a search warrant for that person. And as far as the BATFs laws are concerned, the majority of them are not really that much different from:a law enforcement officials.2) Any citizen that is NOT legally allowed to own firearms(e.g.a semi-automatic pistol, AR-15,and other high capacity firearms) in a military/public place.3) Any person who does not reside at an emergency safe in amilitary/public place.4) ANY person who possesses, sells, or hosts ANY weapon that would, and would NOT, be readily identifiable.	talk
Toxic comments	Okay, then he isn't an idiot, just a dense person. Gute Nacht.	Who likes to write about dicks and other shit on wikipedia.	toxic
	I think the recent editing by all has been brilliant. If I see anything that could be improved I'll speak up here. I would like to see a Christian Theosophy section also. Multiple reliable sources reference it. I'd like to learn more about when the term started. You probably already know. I don't know yet. Thanks for your work (and Gilbert and Donald Richard Sands's work, too.)	I will always be a proud christian and i have a big goal for life you guys are the best and will always be the best for christian history as an individual you are a wonderful person i love you in all your dealings and your work is a big part of my life and you can really add to the history of christian history.	non-toxic
Safeguarding (sent)	Following an outburst, which he said was not physical; he went back to his GP in 2008.	In 2010, he was diagnosed with a chronic form of psychosis which he was unable to identify.	Mental Health Issues
	There is no evidence that he was abusive in nature.	It is rather evidence of a family relationship.	Indicative Behaviour
Safeguarding (pass)	Her General Practitioner with regard to raised blood pressure from February 2010 to shortly before her death. At her last appointment Anne mentioned feeling tired and there was discussion as to whether she was entering early menopause.	There was no indication of her doing any physical activity. In February 2011 there was no physical evidence of any changes in her behaviour or health. A follow-up visit by GP did not produce any further changes.	Contact with Agencies
	From the records it appears that with respect to this information enquires were made with other agencies, but concerns are raised in this report that telephone contact with the Health Visitor did not take place until 14th March 2001. There were no concerns expressed by the Health Visitor, who, apparently, had known the family for some 14 years, although the Health Visitor had 'heard rumours about the father'. There are no further details about the rumours, or what had been done about them.	There is no suggestion that this was linked to any threat to the family, and the issue is under no consideration in relation to these matters. As far as the Health Visitor is concerned the records are all related to his father.The Health Visitor was a highly regarded member of the public and one of the few to be interviewed by police. He was arrested by the MCC in May 1991 and charged with assault and indecent assault.	Reflections

Table 7.4: Generated data using random seed selection and GPT-2 model fine-tuned per label.

Dataset	Original Seed	GPT-2 Generation	Label
20 Newsgroups	Here are the final stats for the CHL for the 1992-1993 season. Central Hockey League Founder and President - Ray Miron Commissioner - Monte Miron Marketing Director - Jim Goodman Information Director - Eric Kolb Each team is owned by the league, with local intrests controlling day to day operations. Working agreements and NHL affiliations are up to each team. Each team has a $100,000 salary cap for 17 total players (16 dress up). Each team is limited to 11 visas, meaning at least 6 players must be Americans (Ignore this rule if you are Tulsa, they only had 2 Americans on the roster during the playoffs). Teams play in Wichita, Tulsa, Oklahoma City, Memphis, Fort Worth, and Dallas. Future expansion plans include Houston, San Antonio, Baton Rouge, Little Rock, El Paso, Albuquerque, Tuscon, and Amarillo.	As of this posting, the top 5 starters are listed, and the bottom 10 are retired, though it is possible that someone may have made a typo. Incorrect numbers could mean that a player has either been out for a game or has been out for a series. If the player was retired before the start of the season, the number will be updated to reflect that, rather than the number of games played. The stats were posted at the end of the regular season, so the last player listed is still listed. In case of a missed player, the players name may not have been updated or not added to the stats. Note that while I am considering new players for this group, there may be players with previous high league numbers. If any players are missing, please let me know.The 1993 draft was not an offensive/defensive draft, but rather an offensive/defensive draft for those players who had played at least 85 games.	recreation
Toxic comments	You swine. You vulgar little maggot. You worthless bag of filth. As they say in Texas. I'll bet you couldn't pour !@#$ out of a boot with instructions on the heel. You are a canker. A sore that won't go away. I would rather kiss a lawyer than be seen with you. You're a putrescent mass, a walking vomit. You are a spineless little worm deserving nothing but the profoundest contempt. You are a jerk, a cad, a weasel. Your life is a monument to stupidity. You are a stench, a revulsion, a big suck on a sour lemon. You are a bleating foal, a curdled staggering mutant dwarf smeared richly with the effluvia and offal accompanying your alleged birth into this world. An insensate, blinking calf, meaningful to nobody, abandoned by the puke-drooling, giggling beasts who sired you and then killed themselves in recognition of what they had done.	Loser if i ever get a chance i will fuck you up and i will make you sick. Hi you are a dick face you fat fuck I am a faggot == What the hell does the Fag of a Woman thing say about a female gag? Fag? You think so? Well, maybe you're just as insecure as me, since you've only been around the Internet for 3 weeks (or so I am certain) and know nothing about feminism, but maybe you are nervous (and possibly paranoid/dangerous/stupid/whatever the fuck you are), as I'm told by many women who have been around the Internet (or who have been around the Internet for 3 weeks or so, since you've only been around the Internet for 3 weeks) and know nothing about women's rights and equal rights. You're not, in my opinion, a feminist either, and you may or may not be.	toxic
Safeguarding (sent)	Wales Probation was initially not made aware of this but, through good liaison with Children Services in Swansea, obtained the minutes of the meeting at a later date, which were then used to inform cognitive behavioural work in supervision and risk management.	Whilst the process of recognising the nature of a potential problem was a lengthy one, the Department of Social Development (DSD) did not make a formal request for any further advice. However, there was a commitment to provide support to the family.	Indicative Circumstances
Safeguarding (pass)	His first conviction for assault was recorded when he was 15 years of age and he subsequently has had numerous episodes of detention/imprisonment along with a range of other penalties for the offences he has committed. In the period within the scope of the review ie between April 2008 and December 2012, he had multiple convictions including those for assaults on police officers (in December 2008 and May 2011), criminal damage (Women's Aid refuge in December 2008) and the assault on his mother (in April 2012), which he was imprisoned for in June 2012. He was known to some agencies as being a perpetrator of Domestic Violence.	In October 2013, a 17-year-old female victim reported that she had been beaten by a male perpetrator at school.He has a history of self-harm, was known to be in a relationship with some male youths who were known to be abusive.In December 2011 he was arrested for assault by police but that was not pursued for many months.During the trial, the Crown agreed to pay a fine of Â£2,000 to the victim of $50,000 for damages caused to her family. He was also given a curfew, which is now a curfew in England for all children aged five to six months.	Indicative Behaviour

Table 7.5: Generated data using max noun-based seed selection and GPT-2 model fine-tuned per label.

Dataset	Original Seed	GPT-2 Generation	Label
20 Newsgroups	Hi netters, Does anyone know have any info on the Ultrastor line of controller? I'm especially interested in the 14F and 34F SCSI controllers. Please email any info ASAP. I'm building a system and that's one of the few components that is missing. Thanks.	Thanks to the folks at VCCI for letting me sell this system to them.I have a VCCI 14F SCSI controller for sale at $15.00 plus shipping (2 copies of the software)and also have a 32MB floppy drive and a 1.5 meg Hard Drive for $150.	for sale
	I am curious if anyone in net-land has spent any time at any of the L'Abri houses throughout the world and what the experience was like, how it affected you, etc. Especially interesting would be experiences at the original L'Abri in Switzerland and personal interactions with Francis and/or Edith Schaeffer.	The discussions were very interesting and both couples had a profound impact on thecommunity. If any of this is going to happen to anyone of you, be it a gay or straight person, please send me a note in advance.Ill post itwhen I get back.Cheers,Kent. In an earlier posting, I noted that there have been several discussions on how we can learn from history by considering the ways in which we have used the scriptures, and how the Mormon tradition has shaped them.	religion
Toxic comments	I wasn't speaking of the Hebrew spellings of Samael verses Samuel but the way the Hebrew name Shemuwel was 'transliterated' by the Latin speaking people of Rome to Samuel which is almost the same as Samael. Obviously because the Jews did say that Samael was the patron of Rome so they must have took it to heart and decided to venerate him instead of choosing a true translieration of Shemuwel. I was being funny when I said Latinese. It is still proper English though. I don't have to speak the Queen's English.	Welcome! Hello and welcome to Wikipedia! Thank you for your contributions. I hope you like the place and decide to stay. Here are some pages that you might find helpful: The five pillars of Wikipedia Tutorial How to edit a page and How to develop articles How to create your first article (using the Article Wizard if you wish) Manual of Style I hope you enjoy editing here and being a Wikipedian! Please sign your name on talk pages using four tildes; this will automatically produce your name and the date.	non-toxic
	Once again, you are taking this way too personally. My apologies if I have offended you. I do not care about edit warring, nor do I think I ever will. This is the internet, learn how to chill out. Not everything here is serious business. If I was blocked from Wikipedia, it would not end the world for me. As it is I have posted a ton of sources validating my claim that there is a relatively popular blog which criticizes Marmaduke on the internet which is read by people in the talk section, as you requested. If people do not believe me after this I will think people have gone defunct. You have still not answered my question about how you have designated authority over me, and I am getting tired of asking you. Why was my posting deleted after several months of being up on the page? Why was a user who is on trail for being a bot allowed to delete my posting? You know, I came here to do one simple thing, not argue on the internet like an idiot. But this whole thing has done terrible damage to my self esteem. I am sorry you do not wish to work with me again, as I think I have eloquently and passionately stated my points, which you have attempted to refute by pointing me to some rules page. I hope that the links I have posted are proof enough, and that I will finally be allowed by the grace of the Wikipedia editors, to post one chunk of information which is true and not inane regarding critcism of Marmaduke. Good fucking lord. I did not think things would be this hard.	please tell me what you want to know when i want to discuss something i cant be blocked just because I didn't know the rules for wikipedia you can edit and if I ask you what the rules for wikipedia are i am going to ask for clarification to be on you little fucking twofan please tell me what you want to know your numbertwofan I was being stupid and i think you have given me the right to post in this thread you cant delete information i was not vandalising but i made it clear as i said i do not intend to vandalise or do i just want to make you understand you can get blocked. I'm just telling you that if i ever want to talk to you again i can and will and that it is to talk to the fact that i did not vandalise this page i am not going to vandalise and i am not going to delete that info you said but i will tell you how to behave. i don't want to be blocked because im not interested you can delete information. i just want to talk to you and im sure you will. i do not have a problem with people who do not agree with me and i am not going to sit here and give away all the info or the rules i asked you to put in. i am sure that by the time im done i will be blocked from this site forever. stop it i was not vandalising you little fucking twofan. Are you gay, or are you a white supremacist? Please, do not attack me with racist remarks. I am not an idiot and you do not deserve to be insulted by racists.	toxic
Safeguarding (sent)	Police and Probation are two of the three Responsible Authorities and could have referred EF to MAPPA.	The other two agencies are assisting the EF in assisting in the assessment of the crime scene.	Reflections
	This also includes PTSD and other disorders carried into adulthood.	PTSD is a mental health disorder that causes emotional distress.	Mental Health Issues
Safeguarding (pass)	This was the only contact North Wales Police had with S.J. Following this incident the officer made telephone contact with the Mental Health Team at Ysbyty Alltwen, Tremadog . The officer was informed that D.J. was known to the team.	This was the only contact North Wales Police had with S.J. Following this incident the officer made telephone contact with the Mental Health Team at Ysbyty Alltwen, Tremadog. The officer was informed that D.J. was known to the team.	Indicative Circumstances
	Throughout the time that Elaine was a client of the Independent Domestic Violence Advocacy (IDVA) Service April 2013 to May 2014 the service was delivered by the charity Advance. Shortly before this DHR commenced the IDVA service had been re-commissioned and the new provider, Hestia, therefore compiled the IMR, with the support of Advance.	The IDVA service is administered in the following way by the Department of Justices Public Safety and Mental Health Unit : in response to calls from victims of domestic violence, to respond to calls from individuals concerned about domestic violence. During the past three years a number of domestic violence calls were made to the DHR for further enquiries relating to sexual assault.	Contact with Agencies

Table 7.6: Generated data using max noun-based seed selection and GPT-2 model fine-tuned per label.

Method	Dataset	Original Seed	GPT-2 Generation	Label
Random	Safeguarding (sent)	On 21/04/05 Mr Z self-referred himself to Options a drug and alcohol counselling service.	The consultation process was followed shortly afterwards and the decision was confirmed.	Contact with Agencies
		When she called again at 17:00, Vasile was calm and she gained the impression that they had made up after the argument.	It has been reported that Vasile was also attacked by her partners boyfriend who also happens to be a taxi driver.	Indicative Circumstances
	Safeguarding (pass)	Twenty minutes later, Damien went to one of Louise's workplaces (Tuition Centre) and left two bags that he said belonged to Louise. Six cannabis plants were in one of the bags. Louise's employers called the police. The manager there said that Damien had been aggressive and reported that he believed that Damien was trying to create problems for Louise.	When police arrived on scene, they found that Damien had taken two of the cannabis plants and cut them off. He said that he had set up the shop at one of her workplaces where he had made no attempt to break into the building or to get into her place of work.A video from the CCTV shows Louise lying in bed.	Indicative Behaviour
		The seventh contact occurred on 4/10/2011 when D.J. telephoned Porthmadog Police Station. The incident log notes that D.J. had been acting in a strange manner which caused concern, that he was alleging the freemasons were after him, however he did not wish for the officer to disclose this to any other officers. The officer requested attendance at the address by a female officer for a welfare check as D.J. did not trust male officers. The officer also commented 'D.J. is 7 saying strange things and is perhaps showing signs of schizophrenia, it was also noted that the officer was concerned that his wife was having problems with his behaviour.	She is apparently having difficulties in sleep and he was concerned about his work for a number of days.On 4/10/2011 the officer reported that D.J. appeared to be suffering from psychosis and that he had not been diagnosed with schizophrenia before. He said that he had been sleeping and had disappeared from the house but that he had also seen friends having sex with D.J.s girlfriend. This was a serious incident which did not result in his arrest. However, D.J. was arrested on the same day and recorded as a witness at the end of the week in the police booking chart.	Mental Health Issues
Nouns	Safeguarding (sent)	It is clear from discussions with Relative 1 that Adult A had a challenging childhood due to her condition, but due to the wishes of Adult A and in order that she was able to lead as normal a life as possible and take part in teenage activities they had nurtured her and kept her in mainstream education.	The parents decided that Adult A should go out as normal as possible without being subjected to further physical abuse or discrimination. This was made clear by the behaviour that Adult A displayed which prompted the police to intervene in January 2011.	Mental Health Issues
		High risk investigations are now reviewed at the daily management meetings which are chaired by senior officers.	However, the department cannot deny that a high proportion of the risk assessments made by senior officers.	Reflections
	Safeguarding (pass)	Neil was born in 1970. He was 45 at the time of his death. His ethnicity is best described as white British (first language English). Neil's mother's details are not known, he is survived by his father. Father and son had not seen each other apart from once in 22 years. Neil had 3 siblings: 'Sarah', 'Jane' and 'Rose' (deceased). Rose's funeral in August 2014 is believed to be the last time Neil spoke to his family. Neil had one male child born in 1994 who was still an infant when Neil was sent to prison.	Neil left his brother after he was released on parole in 2010. His sister and his girlfriend (a partner of his) divorced in 2015, but Neil remained married and had a baby. He lived in a small town on the edge of Birmingham, a family family home. His sister is survived by his sisters-in-law and son-in-law who he had been with for more than 10 years. In 2014 Neil left his father and other relatives to die at their home at 4.30am. Neil died from wounds that left scarring the house and leaving footprints on the floor.	Indicative Circumstances
		Upon receipt of this information had Housing decided not to accommodate EF, there would have been sufficient time prior to his release in July for alternative housing arrangements to have been considered. Linked to this is the wider issue about a lack of suitable accommodation for people who pose a high risk of harm to others. There are no specialist resources for such individuals available to housing. This review questions whether EF's specific needs could have been met in more suitable provision and who has access to such provision, if it exists.	There have been no specific action taken by Housing to ensure EFs mental health and wellbeing can be considered by Health, Disability and the authorities at large. There has also been no clear approach from Health to the Mental Health Officer and the details of the investigation are not known. This review concerns the fact that housing arrangements and other supportive agencies do not appear to address the issue of the adequacy of support or care for those with mental health problems.	Reflections

Table 7.7: Generated data using expert-guided seed selection and GPT-2 model fine-tuned per label.