Identifying cases of drug-harm from electronic patient records using free text analysis

Amiyesh Sahay

Master of Science
Data Science
School of Informatics
University of Edinburgh
2023

Abstract

Scotland is challenged with higher rates of drug harm compared to other developed countries. In this context, the Scottish Ambulance Service (SAS) plays a critical role in responding to drug-related emergencies. While existing methods used by SAS contribute to drug harm identification, an opportunity exists to improve identification using a systematic data-driven approach that leverages textual data within electronic patient records (ePRs). This study aims to understand what classification performance is attainable in the task of classifying drug harm cases and interpret outputs to understand important predictive words and phrases that can be used to also improve the existing rule-based flag that SAS has deployed.

A penalised logistic regression model using n-gram features and TF-IDF weighting was implemented. This served dual purposes: facilitate model interpretation and to provide a baseline for evaluating model performance. The study identified a set of 466 predictive words and phrases associated with drug harm, which will be used combined with expert knowledge from SAS to enhance the existing rules-based flag.

Moreover, this study is the first application of deep learning techniques to ePRs captured in an emergency setting. BERT, DistilBERT, BioBERT, and RoBERTa models were employed in various experiments, exploring hyperparameter tuning in a resource-constrained environment and different text pre-processing approaches that included abbreviation expansion and adding additional features to free-text.

Among the models tested, the highest performing model was BioBERT with additional text processing, which achieved an F1 score of 57.2%, seeing an uplift of 23.4% from the existing flag. Analysis also revealed that the use of diagnostic codes, the current gold standard label, might miss drug harm in complex cases with multiple presenting conditions. This suggests a potentially higher true performance and importantly underscores the model's capability to identify patients that were previously overlooked. This model will be deployed at a national level, enabling SAS to tailor policies more effectively, ensuring the right patients get the right help that they need.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 7583
Date when approval was obtained: 2023-06-23

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Amiyesh Sahay*)

Acknowledgements

I would first like to express my gratitude to my two supervisors, Dr Christopher Lucas and Dr Adam Lloyd. Their invaluable guidance and mentorship throughout this dissertation allowed me to tackle challenges head-on whilst continuously encouraging me to think creatively, push the boundaries of my research, and put into perspective the difference that we were making to help real patients through data.

I would also like to thank my friends, for all of the countless memories we've made together through this journey. What an incredible year it has been.

Finally, and most importantly, I would like to thank my parents and brother for their never-ending support. Your motivation and positivity kept me going through this project and the whole Masters. I couldn't have done it without you.

Table of Contents

Chapter 1

Introduction

1.1 Background

Drug-related harm is a global problem with almost 500 thousand deaths attributed to illicit[1] drug use in 2019 [1, 2]. Scotland has, particularly, faced worse rates than most of the developed world, and has seen rates 4.6 times higher than in 2000[2]. As a result, the Scottish Government declared a National Mission "to reduce drug deaths and improve the lives of those impacted by drugs" [3] with an additional investment of £50 million per year until 2026. A key service that is affected by and is tackling drug harm is the Scottish Ambulance Service (SAS). SAS is the national ambulance service that responds to over 500 thousand emergency callouts per year, providing care to 5.5 million citizens, and is often the frontline care provider for acute drug-related emergencies.

1.2 Identifying drug-harm related patients within SAS

SAS identifies drug-harm patients at three points, two points within the patient care journey and once after-the-fact through data inference. At the point of call, the patient is assigned a **dispatch code** from a selection of 35 different codes (see Appendix A.1) which is deemed to be the main presenting problem whilst triaging over the phone. Once the paramedic arrives at the scene, they also assign a code from the same list, known as the **diagnostic code**. In addition, the paramedics write free text notes to describe the patient's situation at the scene. Any further assessments completed by

[1] use of opioid, amphetamine, cocaine, cannabis, and other drugs

[2] https://www.gov.scot/publications/national-drugs-mission-plan-2022-2026/

[3] https://www.gov.scot/policies/alcohol-and-drugs/national-mission/

hospital clinicians are not accessible to SAS (Figure 1.1).

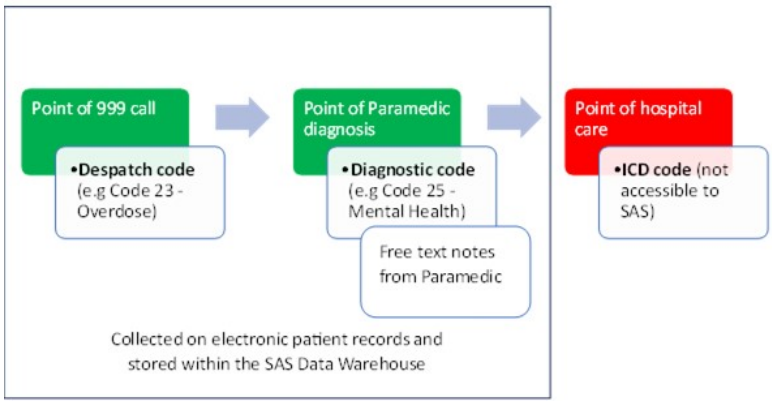


Figure 1.1: Data Capture of patient touch points within SAS. ePR includes dispatch and diagnostic codes, free-text notes and a suite of other salient data points. Stored within SAS DataWarehouse.

For inferring drug-harm from the data in an electronic patient record (ePR), SAS have a rules-based flag called the non-fatal overdose (NFOD) flag. This flag identifies drug harm if the paramedic has checked a box signalling: 'naloxone was given', or the 'substance affecting condition' is 'opioids' or 'street benzodiazepine', or if any of the four following words are present in the free-text; 'Naloxone', 'Methadone', 'Narcan' and 'Heroin'. Investigating the three points is crucial to understanding the improvements to be made.

1.3 Problem statement and hypothesis

Despite the service collecting a vast amount of data within the ePRs, there is currently an absence of an effective systematic and data-driven approach to leverage the true value of text data and gain detailed insights to help identify the right patients for support. Across the three points to identify drug harm, there need to be analyses to understand potential improvements to be made.

The only previous empirical application within SAS, that investigated this type of ePR was conducted by Manca et al. [3] who demonstrated that by using machine learning techniques, on a similar dataset, it is possible to identify patterns within the free text and drive more precise predictions of alcohol-related harm in comparison to human classification. This showed a lot of promise and potential for leveraging information from the text and raises the hypothesis that there are, similarly, finer nuances between

drug-harm and non drug-harm text that can be identified through the use of more advanced natural language processing techniques.

1.4 Research Aims and Questions

This project aims to address the opportunity to better identify cases that are related to drug harm by analysing free-text data. It proceeds incrementally, first understanding the performance of the current state approach within SAS, then establishing a baseline model for improved classification. The goal subsequently is to use deep learning techniques, specifically BERT models, to develop a systematic, data-driven approach to identifying patterns in the free text, and classifying cases into drug-related harm or not. The study details explainable additions to the current rules-based approach, builds classification models using state-of-the-art techniques, and provides directions for future research. Therefore, two research questions are addressed:

1. (RQ1) What performance can be achieved in detecting likely cases of drug harm, and what models and features facilitate accurate classification?

2. (RQ2) What keywords and phrases in the free text are indicative of drug-harm-related cases?

1.5 Research objectives

RQ1 and RQ2 have distinct objectives. RQ1 seeks to transform SAS's capabilities, using state-of-the-art classification techniques, to drive an increase in precision and recall. This will mean that fewer patients that are impacted by drug harm are missed by the service (recall), whilst limiting an increase in inefficient use of resources (precision) in delivering proactive support.

RQ2 objective is to provide interpretable insights into important words and phrases to improve SAS's current NFOD flag. Since the NFOD flag uses a dictionary look-up to find words, the intricate semantic context and nuances found by BERT cannot be implemented, therefore the stakeholder's requirement is to find additional words or phrases (that do not depend on surrounding context). Consequently, a deliberate choice was made to select a simpler machine learning model, penalised logistic regression, due to its inherent interpretability. This provides actionable insights for SAS's team to improve the current rules-based flag whilst meeting the requirements to explain the

rationale of the identifier when for example, reporting drug-harm-related figures or freedom of information requests.

1.6 Expected contributions

In this study, contributions are made across two layers: academic advancement and clinical enhancement.

Academic advancement There has been very limited work done on text classification on ePRs captured in an emergency setting. Building on the previous alcohol-related study [3] conducted in SAS which used a random forest algorithm as their classifier, this study marks the first application of deep-learning techniques, which is commonly applied to clinical in-patient medical reports, to this type of data. By doing so, the upfront costs and time burden of manually hand-crafting features are removed, instead allowing the model to learn directly from the vast amounts of data that SAS captures. This lowers the barriers and provides a foundation for further work to be done to fully leverage the potential of emergency medical data within SAS.

Clinical enhancement Additionally, the investigation seeks to improve the current practices that SAS employs to identify drug harm, better enabling the service to have a more comprehensive understanding and facilitate investigations at a more granular level on cases that would previously not have been identified.

The model will be deployed at a national level, providing SAS and broader healthcare professionals with the latest cutting-edge deep learning techniques to proactively improve patient care and to make more informed policy decisions using new actionable insights.

Chapter 2

Relevant Work

This chapter covers the background knowledge required to understand the approach and findings of this project. Section 2.1 provides a landscape of text classification techniques, and 2.2 discusses applications in the medical space. Section 2.3 then addresses further considerations specific to the medical domain that will need to be addressed in this project. Sections 2.4 and 2.5 delve into BERT and its model variants that could be applicable to clinical text, and finally section 2.6 gives a spotlight on interpretability.

2.1 Text classification landscape

The landscape of text classifiers covers rule-based, traditional statistical and machine learning classifiers, and then more recently, deep learning based classifiers. Rule-based examples will be briefly covered in the next section but given most recent literature focuses on the latter technologies, the focus will be on those in this section.

Traditional statistical and machine learning classifiers

The usual approach is two-step, beginning with feature extraction, then followed by classifier [4]. Features are typically extracted using frequency based techniques such as Bag of Words [5], n-grams [5] or word embeddings such as word2vec [6] to represent text data as numerical features, with different types of feature weighting such as term frequency inverse document frequency (TF-IDF) [7] to give stronger importance to rare predictive words. Classifiers then use these features to make decisions. Algorithms such as Naive Bayes [8], support vector machines (SVMs) [9], logistic regression, and random forests [10] have been widely used for text classification. Naive Bayes

takes frequency counts of features and uses Bayes' rule to calculate probabilities of being in a class. Whilst is it known for its simple implementation, it assumes that each feature is independent and hence struggles to capture relationships between words. Logistic regression is a linear classifier that also calculates probabilities of belonging to a class by assigning weights to different features. It has been a popular choice given the ease of implementation and ability to interpret using the coefficients of the features. However, in high dimensional spaces can be prone to overfitting. An effective solution is often used, such as in the study by Genkin et al. [11], where an L1 lasso regulariser is used to promote feature sparsity, hence improving generalisability. Moreover, SVMs, which find hyperplanes that best separate the two classes, are also shown to be very effective in high dimensional spaces, where there are more dimensions than observations, and differentiate from the previous approaches by also working well on non-linear classification tasks. A challenge however is that they are computationally expensive on large datasets. Similarly, random forests which take the outputs from a combination of decision trees to make a classification, are a popular choice on high dimensional tasks. It benefits from the ability to understand interactions between features, manages sparsity, and is less prone to overfitting since it takes in the input over several trees before making a decision [12].

One of the key challenges associated with the traditional classifier models is the need for feature engineering. Expert knowledge is required to create useful features which can be highly time-consuming and costly [13], and limits the ability of the model to learn hidden nuances and relationships between words from the training data.

Recent deep learning approaches

Without the requirement to hand-engineer features, deep learning models are trained on large datasets with a focus on learning the relationships within the data to make effective classification models. This reduces the requirement for domain experts and reduces the time taken to build the model and implement and reduce costs [13]. In the absence of funding or significant time, this feature of deep learning becomes very useful.

Within deep-learning language models, the best-in-class broadly can be split into traditional neural networks and transformers [14]. Traditional neural networks include RNNs and LSTMs. Recurring neural networks (RNNs) work by processing words one by one through individual neurons, where the hidden state of the previous word serves

as the context for the subsequent word. This sequential approach allows the model to pick up short-range contextual relationships effectively. However, they struggle with long-range dependencies due to the vanishing gradients problem; where context from older words diminishes as is passed through a sequence (due to small gradients). This is where LSTMs then become beneficial. LSTMs introduce a hidden layer of memory blocks into the architecture with a trainable linear memory cell hence mitigating any vanishing gradients and allowing the model to learn longer-range dependencies in the text better.

Transformers benefit from the power of the attention mechanism [15], which assigns weights from each word to every other word in the sequence based on their significance. This mechanism makes it possible to capture long-range contextual dependencies across the full sequence, and importantly, simultaneously through parallelisation as opposed to traditional neural networks using sequential word-by-word processing. This capability enables the training of significantly larger language models on GPUs faster [15]. Using the transformer technology, the pioneering BERT model [16] was developed (discussed in detail in section 2.4). BERT continues to be the state-of-the-art embedding model across several classification datasets [14].

2.2 Free text analysis in a medical setting

Electronic medical files hold a vast amount of free text, providing valuable information about patients. As a result, extracting information from this text has been an active area of research [17, 18, 19]. Effective text classification can help healthcare professionals to gain a granular understanding of patients, including symptoms, diseases, and medications, leading to improved patient care. Rules-based, machine-learning, and deep-learning approaches have been tested extensively in this space.

Rules-based approaches such as dictionary look-ups and regular expressions have traditionally shown good scores, for example in a review by Spasic et al. [20], using this approach for the identification of cancer-related words in ePRs achieved F-scores between 80% and 90%. Yang et al. [21] took a hybrid approach, combining rules-based and machine learning. Using hospital discharge summaries that contained multiple categories, including “Diagnosis”, “Past or Present History of Illness”, and “Medication/Disposition”, they were able to effectively predict obesity and 15 related diseases with a macro F-score of 81%. Moreover, Bates et al. [22] achieved an impressive F1 score of 93.5% by using an SVM model on radiology reports to identify patients who

have experienced falls. Heo et al. [23] employed several deep learning models such as convolutional neural networks (CNN) and long short-term memory (LSTM) to predict poor outcomes from stroke patients using brain MRI text reports.

Comparing the literature in this space, whilst there have been several applications using different medical reports, a holistic patient understanding such as past medical history is often used as an important feature for making classifications. This level of context is generally not available in ePRs using emergency text, and therefore, the same application of models on these reports could see a drop in model performance. Moreover, compared to general text, medical text has a number of unique properties such as medical jargon, abbreviations, and poorer grammar [18, 24], and therefore focus has predominantly been on text pre-processing to design informative features.

As discussed in the previous section, BERT has consistently been shown to produce state-of-the-art results, but Lee et al. [25] showed that it struggles to generalise as well to the biomedical space. Mascio et al. [26] tested a number of algorithms on medical text classification including BERT, BioBERT and found that without any additional customisation, the general domain and domain-specific BioBERT model outperformed all of the other models - 93.4% F1 score for BioBERT and 91.5% for BERT vs 88.4% for the next best performing model - bidirectional LSTM. They argued that training on domain-specific text improves ability to clinical text classification. Some domain-specific BERT models include; BioBERT, ClinicalBERT, MedBERT are likely to perform better by effectively picking up longer-range dependencies by recognising domain-specific terminologies [25, 27, 28].

Applications on ePRs with emergency medical text

There has been very limited work on ePRs using emergency medical text. Prieto et al. [29] looked at the application of machine learning techniques on emergency text to address the challenges of misclassifying cases with naloxone as opioid use cases. They applied four machine learning models; random forest, k-nearest neighbours, support vector machines, and L1-regularized logistic regression to classify cases associated with opioid and heroin misuse. They found that the L1 regularised logistic regression model performed the best, and improved precision by 30.3% from their rules-based flag. Manca et al. [3] conducted a study looking at the burden of alcohol-related cases on the ambulance service, which demonstrated the richness in the text to effectively identify alcohol-related cases through the use of random forest algorithm on emergency medical reports. Senior paramedics reviewed over 5k ePRFs and pulled words and phrases

that were alcohol-related. These words and the existing alcohol flag were then used as features in a random forest model, which showed a large improvement in sensitivity (0.942 for RF vs 0.380 for the existing flag). No study has looked at deep learning techniques on this type of data. By applying deep learning techniques, the requirement in Manca's study for a experienced paramedic create features is removed, reducing the costs, and time burden to go from model build to implementation.

2.3 Additional considerations with medical data

2.3.1 Privacy and anonymisation

Ensuring the privacy of patients is respected whilst reducing the risk of potential bias in training any deep learning model [30, 31] is paramount and core to this study. Given the unstructured nature of the free text, there is a risk that identifiable information such as patient names may be contained in this text.

A common approach to automatically anonymse data is token level classification, using named-entity-recognition (NER) [32]. NER is the task identifying words or phrases that are associated with names, places, or organisations [33]. Garcia et al. [34] tested both spaCy and BERT on the detection of sensitive words and classification of the type of word (e.g. name, age, location). Under detection, they showed BERT outperformed across each dataset with an F1 score of 96.5% vs 95.1% for spaCy. They also suggested different forms of anonymisation of sensitive data, through complete removal, replacement or obfuscation [35]. For example; "64-year-old patient operated on a hernia on the 12/01/2016 by Dr Lopez" can be replaced with "XXXX patient operated on a hernia on the XXXX by XXXX", "[AGE] patient operated on a hernia on the [DATE] by [NAME]", or "59-year-old patient operated on a hernia on the 05/06/2019 by Dr Sancho" [1]. Interestingly, Dayanik and Pado [36] found that masking names in training removes personally identifiable bias but also improves model performance in out-of-domain settings.

2.3.2 Imbalanced data

Class imbalance typically leads to poorer classifier model performance [37]. Usually, higher levels of misclassification are observed in the smaller-sized class, also known

[1] translated from Spanish to English from [34]

as the minority class. To address imbalance, researchers often use undersampling or oversampling techniques [18]. Undersampling is a popular and efficient method to address this issue, by taking only a subset of the major class (cases that are not related to drug-harm) [38]. Oversampling techniques, such as SMOTE, synthetically create instances of the minority class, thereby artificially increasing the size of the set [39].

However, there are challenges towards both approaches, where undersampling can lead to important information being removed and over-sampling can lead to overfitting [40]. Sullivan et al. [41] assessed text classification using both undersampling and oversampling for detecting misdiagnosis of Epilepsy. They found that their best-performing model used an undersampled dataset to achieve an F1 measure of 71.4% compared to 64.7% using oversampling. Similarly, Afzal et al. [40] compared the two techniques on two imbalanced datasets for classifying cases of acute renal failure and hepatobiliary disease and found both approaches drove improvements, but undersampling performed slightly better. However, Garcia et al.[42] found in cases of severe imbalance, oversampling performed better than undersampling, likely due to the level of discriminatory information loss from the majority dataset. This suggests that there is not only one 'best approach' with choices dependant on the dataset being used, such as the level of imbalance and variability of text within each class.

Alongside rebalancing, model choice also impacts performance on imbalanced data. Lu et al. [43] applied different deep learning models, including, CNN, Transformer encoder, and BERT on medical text data with varying levels of imbalance. They saw transformer encoders were the most resilient to varying levels of imbalance and surprisingly BERT did not perform as well. However, they observed the use of domain-specific embedding BioWordVec had a positive impact on performance, hence could suggest that domain-specific BERT such as BioBERT could also show more resilience under imbalance and therefore is interesting to explore.

2.3.3 Abbreviations

Abbreviations and acronyms can reach up to 50% of words in clinical text, compared to less than 1% in general text [44]. Surfacing more words that are human and machine understandable is likely to help the model make better classifications. As a result, a number of comprehensive datasets have been produced for medical abbreviations [45, 44]. A challenge with abbreviations, however, is that each abbreviation may have several potential expansions, hence correct expansion is a critical pre-processing task in

order to help improve model performance. For example, The Medical Abbreviation and Acronym Meta-Inventory [44] contains over a hundred thousand abbreviations, within which there are several expansions for each abbreviation, such as 'OD' expands to 34 full forms such as 'overdose', 'optimal dose', and 'occupational dermatitis'. To identify the most likely expansion, Pakhomov et al. [46] used cosine similarity to measure the similarity between training and test context vectors. The vector corresponding to the largest cosine (greatest similarity) would then be picked to represent the expansion of the acronym. Furthermore, Liu et al. [47] explored the use of word embeddings measuring cosine similarity but then also combined with a rating score, based on the popularity of the word, and saw an accuracy of 82%. Whilst these approaches perform well, they fall slightly behind domain expert accuracy of c.90%.

2.4 A focus on BERT

BERT (Bidirectional Encoder Representations from Transformers) [16] is a deep contextualised language model. Its architecture is structured on a stack of 12 layers of transformer models, first introduced by Vaswani [15]. Each layer uses a mechanism called self-attention to assign a weight between each of the words with every other word in the sequence (both directions), hence capturing long-range contextual relationships within the text. BERT benefits from transfer learning [48], where it is first trained on a source task, and then fine-tuned onto a specific task. During pre-training, BERT uses an unsupervised masked learning model approach and next sentence prediction [16] using a large unstructured dataset of 3.3bn words from BooksCorpus and Wikipedia. This enables BERT to learn language patterns. During fine-tuning, all of the parameters are tuned together to learn further intricacies of the language used in a specific task.

2.5 Variants of BERT

Following the success of BERT on a range of NLP tasks, both in general domain and in a clinical setting [16, 14], a number of variants have been developed with different techniques to improve performance, such as looking at domain-specific cases, improving BERT model through better training, or different size models. With benefits and challenges with each model, Minaee et al. [14] provide a helpful framework for model selection, with suggestions to look at domain adaption, model design, availability of training class labels, and consideration for real-world feasibility. As a result, a range

of general models; BERT, RoBERTa, and DistilBERT, and a domain-specific model; BioBERT are implemented in this study.

2.5.1 RoBERTa

RoBERTa (Robustly optimised BERT approach) [49] improves on the original BERT model, by increasing training time over significantly more data, removing the next sentence prediction objective and adapting the pre-training process by dynamically changing the masking pattern applied on training. Liu et al [49] showed the RoBERTa model to outperform the BERT model in all tasks, importantly on the SST task, which looks at text classification, by delta of +2.9% accuracy in comparison to BERT.

2.5.2 BioBERT

BioBERT [25] uses the same architecture as BERT and takes the weights initialised from the BERT model [16], but then differentiates itself in the data that it is pre-trained on. BioBERT pre-trains on biomedical corpora using 4.5B words from PubMed Abstracts and 13.5B words from PMC full-text articles. This has been done since vast amounts of biomedical text are different to general text, and therefore generally BERT would see a performance drop-off when used on this domain specific text. Comparing performance of different text classification approaches on electronic health records, Mascio et al. [26] showed BioBERT consistently performed better than BERT, for example achieving a macro F1 score of 93.4% vs 91.5% respectively on a task to classify if the disease is affirmed or negated.

Other domain-specific models have been developed, for example, ClinicalBERT [27], which takes BERT and BioBERT models fine-tuned on clinical notes. Interestingly, Turchin et al. [50] compared the performance of BioBERT and ClinicalBERT on different tasks such as to classify usage of tobacco in the past, and did not find either model consistently performing better than the other. Combined with the explained differences between emergency reports and clinical text, later outlined in section 3.1.3, for this study, BioBERT has been chosen for the domain-specific model.

2.5.3 DistilBERT

DistilBERT uses knowledge distillation [51] to reduce the size of the BERT model by 40% whilst still retaining 98% of classification performance [52]. For example,

on SST-2, which is a sentiment classification task on movie reviews [53], DistilBERT demonstrated an accuracy score of 91.3% [52] compared to 93.5% [16] for BERT-base. The reduction in the size of the model means inference time increases by 60%. The trade-off between training time and performance of using DistilBERT is interesting to explore if a similar delta in performance is seen within a clinical medical setting as it does in a general setting. These findings will be used to provide insights for approaches to future text analysis tasks where there are hardware limitations stopping fast training, within SAS research.

2.6 Model interpretability

Deep learning models, due to their 'black-box' nature [54], often lack transparency, making interpretability a challenge. Explainable AI has become an increasingly popular area of research. Different techniques have been explored, such as investigating BERT attentions [55] and deployment of libraries such as Captum [56] that use different types of algorithms such as layer importance algorithms improve the ability to interpret. However, debates persist about the effectiveness of these techniques, and is often argued, such as in the studies by Rudin [57, 58], that in high-stakes areas such as healthcare, the focus should be given to building interpretable models rather than explainable AI, since the explanations derived are often wrong.

In contrast, traditional text classification models offer simpler interpretability through feature selection. Among these, logistic regression stands out for its simplicity and human interpretability. Since the output is a weighted sum of features, the magnitude and sign of the coefficients can be interpreted to understand feature importance. The benefit of feature selection such as on n-grams is that these are directly interpretable by humans by investigating what is the most important for the model [59]. However, it must be noted that feature selections show features *correlated* with the outcome, not *causal* [59], hence expert knowledge must still be applied to interpret the outputs of a model.

In this study, providing interpretability of the model will support SAS in improving their rules based flag in a systematic and transparent way so they can continue to use their flag for governance purposes. As a result, a context-free model is required since a rules-based flag that uses a dictionary to look up words or phrases need not depend on words in close proximity. Hence, the implementation of the BERT model which is a contextual model is not used here.

Chapter 3

Methodology

3.1 SAS Drug-harm dataset

The dataset used in this study contains 47k electronic patient records (ePRs) that have been captured by SAS over full year 2022. As outlined by Figure 1.1, there are two points in the patient's journey with SAS that information regarding drug harm is recorded; at point of call where a **dispatch code** is provided and at point of paramedic diagnosis where a **diagnostic code** is added. In addition to these codes, there are other fields of data that are provided within the dataset that are captured in table 3.1 below.

Cohort	Call Number	Date	Time
Call despatch code	Diagnostic code	Call colour	NFOD flag
Naloxone mentioned	Heroin mentioned	Additional comments	Postcode
Receiving hospital			

Table 3.1: Fields provided in SAS dataset. Definitions in table A.1

As advised by SAS clinicians and in line with previous applications of supervised learning text classification tasks where clinical codes are used as class labels [19], the golden truth of if a case is related to drug harm or not in this study is determined by the **diagnostic code** 23 (see figure A.1). Drug harm & OD is used interchangeably.

3.1.1 Selecting final dataset for study

4.1% of all emergency callouts are related to drug harm. To construct the training and test dataset for use in this study, two considerations were taken into account. Firstly,

managing the dataset size with consideration to training time, and secondly, class balance was prioritised to improve classifier performance [38, 18].

As discussed in section 2.3.2, undersampling techniques and oversampling techniques perform better in different situations. Due to the practical limitations with training time for fine-tuning BERT, the training dataset has been kept at a sample size of 10k of the 47k cases. Given the constrained dataset size, with cases already being removed from each cohort, the undersampling approach was deemed more suitable and was selected. Specifically, 5k of OD-related cases (by diagnostic code) have been randomly selected, proportionate to the distribution in cohorts 1 and 3, and 5k of not-OD related cases randomly selected, proportionate to cohorts 2 and 4. Whilst there is a risk for information loss from undersampling, the dataset is deemed suitably large (at this stage) to give sufficient information for the model to effectively learn from the positive and negative classes.

	Description	**Original**	**Training**	**Test**
Cohort 1	Dispatch: OD, Diagnostic: OD	9.5k	2.2k	270
Cohort 2	Dispatch: OD, Diagnostic: No OD	5.0k	50	110
Cohort 3	Dispatch: No OD, Diagnostic: OD	12.3k	2.8k	340
Cohort 4	Dispatch: No OD, Diagnostic: No OD	20.0k*	4.9k	14.0k

Table 3.2: SAS dataset cohort split. Training was rebalanced to 50:50 split: Cohort 1 & 3 = 5k, 2 & 4 = 5k. Test kept original cohort split. *from a subset of c.500k no-OD cases.

Model selection was performed using training set. The remaining data was used (to ensure cohort 1 & 3 were large enough) as the test dataset with proportions matching the original set. Note, that a limitation of the dataset provided was that cohort 4 was only sampled from May & December due to current SAS infrastructure limitations. Therefore an assumption was required that cases did not vary by condition across months.

3.1.2 De-identifying the Dataset

The data was first anonymised before training to avoid potential bias and respect the privacy of patients. Postcodes were condensed using regular expressions to retain area and district information. Automatic anonymisation was chosen due to the impracticality

of manually deidentifying free text in 10k ePRs. This was an accepted approach by SAS and agreed in the Ethics Approval since the data used remained within the SAS environment for this study. García-Pablos et al. [34] showed that a general BERT-based model performed well in de-identifying and did not require any domain-specific engineering, as a result, BERT_NER[1] was used. BERT_NER, fine-tuned cased model on the CoNLL-2003 Named Entity Recognition dataset [60] - a commonly used public dataset for named-entity-recognition, classified each token into one of 4 classes: person, organisation, location, or other (see table A.3 for full breakdown).

Within the 10k rows of data, 5.9k names, 2.7k locations, and 14.8k organisations were identified. Names were masked with "PERSON" rather than obfuscation in line with findings from [36] (see section 2.3.1). For example, "Dr Smith" was replaced to "Dr PERSON arranged a home visit". Upon agreement with SAS, locations and organisations were not masked given it was not personally identifiable information and since the data remained within the secure SAS environment.

3.1.3 A note on the differences between SAS ePR dataset and hospital ePRs

Here it is important to note, that whilst certain similarities exist between medical notes captured in a clinical setting by healthcare professionals, and emergency notes documented by paramedics, they generally serve different purposes and therefore have different characteristics. Electronic patient records used in this study are designed to specifically capture a comprehensive description of the event that paramedics are attending. Unlike clinical notes, which encompass much broader patient histories, emergency medical notes generally do not capture this. Moreover, emergency medical reports are generally more formulaic and quantitative, with specific abbreviations to the service, and designed to provide relevant and important information for handover to nurses and doctors in Accident & Emergency departments. There is often less time to write detailed documentation in these time-critical callouts as compared to in a clinical setting [61]. These differences could pose additional challenges in classification due to limited context. For example, a patient with a history of specific clinical events, such as mental health related, could be a likely predictor of future overdose [62].

[1] https://huggingface.co/dslim/bert-base-NER

3.2 Methods

This section introduces the two methods used to classify the ePRs; logistic regression and BERT model (including variants).

3.2.1 Logistic Regression with L1 regularisation

The logistic regression model is used as a baseline for RQ1, for feature selection for adding context for the BERT model, and for interpreting keywords to answer RQ2.

Pre-processing

Using the dataset with 10k ePRs, 90% were randomly selected for training. To answer RQ2, the remaining 1k ePRs were used as the test set. The choice was made as the focus was on interpretability, rather than model generalisation, hence the aim was to assess the model's ability to distinguish between drug-harm and non-drug harm cases without being biased by class imbalance. However, for RQ1, the performance evaluation was on the unbalanced test set to assess how well the model generalises to new data (table 3.2). The 9k training data was then further split into train and validation data by using 10-fold cross-validation to tune the L1 regularisation parameter, λ.

The free text was pre-processed, which included the removal of punctuation, changing all characters to lowercase, removal of stop words (such as a, the, and) but keeping negation words (e.g. not), and reducing each word to their lemma (dictionary form) [18, 5, 63]. Example of preprocessing: "Called to Patient who has reported has having taken a overdose." to "call patient report have take overdose".

Feature extraction and weighting

To extract features from the text, the Bag Of Words (BoW) method [5], where the number of times the features (e.g. words) have appeared in the corpus is counted, was selected due to its simple implementation and ability to interpret after. This was in line with the common approach in clinical text classification using machine learning as explained in a systematic literature review by Mujtaba et al. [18]. Unigrams, bigrams and trigrams [64, 65] were selected as features since it helps to understand both individual words that are predictive but also areas where a combination of features is more informative than the word on its own, e.g. is "take overdose" more predictive than

"take". In several studies [66, 67, 68], using bigrams and trigrams is shown to drive an uplift in performance than just using unigrams.

Term Frequency-Inverse Document Frequency (TF-IDF) [7] transformation was then applied creating an N x M matrix X, where N = number of ePRs and M = number of n-grams, for defining the feature weighting. TF-IDF calculates the frequency of words within the document (term frequency) and then how rare a word is in the entire corpus (inverse document frequency), and is defined as:

$$TF\text{-}IDF(n,m) = tf_{n,m} \cdot \log\left(\frac{N}{df_m}\right) \tag{3.1}$$

where $tf_{n,m}$ is term frequency of n-gram m in ePR document n, df_m is number of ePRs that have n-gram m, and N is total number of ePRs in set.

The benefit of this is that rarer words that are often found in a group of similar documents will be given a higher TF-IDF score, and common words that are found in all documents will then be given a lower score. This is useful to address RQ2 to identify specific keywords that are more indicative of drug harm.

Training and evaluating

A logistic regression model was then applied using an L1 Lasso regulariser (reasons highlighted below). The objective when fitting the logistic regression model is to minimise the corresponding loss function (to get predictions close to the true value). The loss function for logistic regression with an L1 regulariser [69] is defined as:

$$min\left(\sum_{(x,y)\in S}(-ylog(y_{pred}) - (1-y)log(1-y_{pred})) + \lambda\sum_{i=1}^{N}||\beta_i||_1\right) \tag{3.2}$$

Where $(x,y) \in S$ represents the x values and y labels in training data, y is the true label, and y_{pred} is the probability between [0,1], calculated by $y_{pred} = \frac{1}{1+e^{-\Sigma_{i=1}^{N}\beta_i x_i}}$. The L1 regulariser $\lambda\sum_{i=1}^{N}||\beta_i||_1$ adds the absolute value of the coefficient weights to the loss function, penalising large coefficients, and hence is used to shrink many regression coefficients to zero. This promotes sparsity [70, 71] and helps to understand which features are most significant. The choice of λ was determined by conducting a grid search of different values and selecting λ corresponding to the highest 10-fold cross-validation F1-score. Then using the optimal λ regulariser, the model is trained on the training set, coefficient interpreted and performance evaluated on the test set.

3.2.2 BERT model and variants

As discussed in section 2.5, several variants of BERT have been developed and demonstrate excellent results in different scenarios. Having established a baseline using logistic regression for RQ1, the study now tests whether applying these models in this context sees similar generalisation as observed in other general and clinical scenarios [14, 28]. Since this study is the first application of BERT on ePR that contains emergency medical data, due to the additional challenges outlined in section 3.1.3, this study has tested general domain BERT models; BERT, DistilBERT, RoBERTa, and a domain-specific model BioBERT. This investigation provides empirical evidence into how these models, commonly used on hospital ePRs, adapt to emergency text.

BERT approach

There are two approaches to using the BERT model for a classification task.

1. **Fine-tuning:** A final fully connected linear layer is added on top of the BERT architecture. The [CLS] contextualised vector is mapped to the labels in the SAS drug-harm dataset by **tuning all of the parameters end-to-end**.

2. **Feature-based: Freezing the pre-trained BERT model parameters**, the model first converts the sentences into contextualised vectors. The corresponding vector representing [CLS] token is then taken from the last few hidden layers of the model and then used in a further classifier model, along with other features [72].

Whilst performance is broadly similar, Devlin et al. [16] demonstrated that the fine-tuning approach produces better results on average, for example on a NER task, the fine-tuning approach had an uplift in the F1 score of 0.3% as compared to the best feature-based approach. This is also supported by findings by Peters et al. [73] that showed an 0.5% accuracy uplift on a sentiment analysis task on the SST-2 dataset [53]. As a result, the fine-tuning approach was selected for this study.

BERT implementation

The implementation for each variant is similar, so the implementation of BERT is discussed here. Free text is initially transformed before using BERT. Each ePR text document is tokenised and converted into a numerical vector through the summing of three embedding layers; token, position, and segment embedding. Token embedding uses WordPiece embeddings [74] which breaks words into tokens using a vocabulary of

around 30k words. Two special types of tokens are also added; [CLS], which goes at the beginning of the sequence of text and is critical in classification tasks to help BERT understand the overall context of the document, and [SEP], which goes at the end of each sentence to indicate when a segment has finished. Tokens are then mapped to unique IDs based on the vocabulary. WordPiece handles out-of-vocabulary words well by breaking them down into subwords. The segment embedding and position embedding are used to indicate the segment, i.e. which sentence it is in (in a classification task is always segment 1), and position (what number the word is) for each token.

BERT's input length can go up to 512 tokens. However, most ePR documents (figure 3.1) are under 256 tokens - there are 11% ePR documents that are longer than 256 tokens (10% of drug-harm ePRs and 13% of non-drug-harm). Due to the self-attention mechanism in BERT, where there is a weight attached between each word and every other word in the sequence of text, computational time grows disproportionately (per layer $O(dn^2)$ where d is vector dimension and n is token length) with the length of the free text [16]. Therefore to balance the trade-off between training time and potential loss of context, the max length was set at 256 tokens.

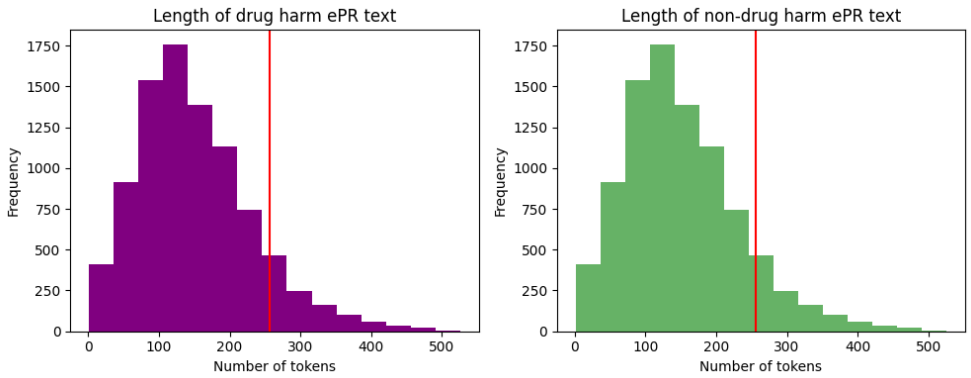


Figure 3.1: Token length of ePRs in dataset

Any ePRs longer than 256 tokens are truncated from the right, as from manual inspection, paramedics typically start by writing the most critical information. For ePRs below 256 tokens, [PAD] tokens appended from the right, which can be thought of as empty tokens, to ensure all vectors are the same length. An attention mask is then used to identify padding tokens (0) vs real tokens (1). See example in appendix B.1.

The input layer is then passed through the BERT model (details in section 2.4), and the outputs are contextualised vector representations of each token. For classification, only the [CLS] token contextual representation is used, which can be thought of as the contextualised summary of the whole ePR document. A fully connected linear layer is

attached to the model which maps the [768,1] vector representation of [CLS] to [2,1] logits (drug-harm or no drug-harm). A softmax function then converts the logits into probabilities using:

$$softmax(x_i) = \frac{e^{x_i}}{\Sigma_{k=1}^{2} e^{x_k}} \tag{3.3}$$

All of the parameters are tuned end-to-end against the training labels via backpropagation with AdamW optimiser [75], minimising cross-entropy loss (equation 3.4). The implementation of the tokeniser and model was in Python, using the HuggingFace Transformers Library [2].

3.3 Additional feature selection

To use fine-tuning end-to-end, features that provide the most information gain were first identified and then appended as free text, e.g. “location: Edinburgh”. Feature selection was required so unnecessary noise was not added and given the token length constraint. With no ’one size fits all’ approach for feature selection [70], an ensemble approach to feature selection was taken (wrapper method and embedded method). An aggregation of multiple selection approaches was used as it has been shown to increase the stability of final model performance results [76].

3.3.1 Wrapper method

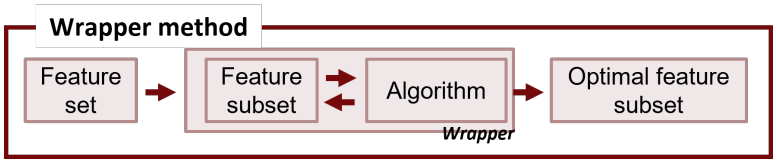


Figure 3.2: Wrapper method [77]

Using backward selection, the process starts off with the full feature set and **iteratively** removes the least important features which is determined from a classifier algorithm (in the wrapper box) until a fixed number of features is reached, yielding the optimal feature subset. To overcome the time-intensive nature of fine-tuning BERT with feature combinations, a faster random forest tree algorithm was used as a proxy classifier, inspired by Manca et al. [3], who, using a random forest, achieved high classification performance on a similar SAS dataset.

[2]https://huggingface.co/docs/transformers/index

A challenge with the wrapper method is that, while the Random Forest model can capture feature interactions, the backward selection may still overlook combinations of features that collectively contribute to improved performance. This could lead to a sub-optimal set of features. For example, removing the 'day of the week' while keeping 'hour' could significantly reduce the importance of 'hour', even though in the context of, for example, Friday, 2am is highly informative. This reinforces the value of taking the ensemble approach.

3.3.2 Embedding method

As discussed in section 3.2.1, using the L1 Lasso regression is an effective method to shrink many regression coefficients to zero and promote sparsity [70, 71], driving feature selection. The size of the regulariser λ was carefully selected, where the stronger the regulariser, the more coefficients would be forced to zero. A target number of features was set at 18, and then the value of λ was amended through trial and error until this target was achieved. Furthermore, k-fold validation (see section 3.4.3) was employed, where k = 10 to ensure stability and confidence in the results.

3.4 Evaluation metrics

3.4.1 Loss function

To evaluate the different hyperparameter combinations, the aim was to minimise cross-entropy loss [78], defined by:

$$\hat{H}(P,Q) = -\Sigma_{i=1}^{N} p_i \cdot log(q_i) \tag{3.4}$$

Where P is the probability distribution of predictions, and Q is the distribution of the true labels. $\hat{H}$ quantifies the similarity between P and Q. A higher value of $\hat{H}$ indicates greater dissimilarity between the predictions and true labels, signifying poorer predictions. This is a commonly used loss function in BERT fine-tuning tasks [16, 79, 80].

3.4.2 Classification metrics

A confusion matrix (table 3.3) compares the model predictions and true labels into a grid. For model selection, loss and metrics in table 3.4 are reported. The model with the lowest loss and higher F1 score on the balanced validation dataset was selected.

	True positive class	**True negative class**
Predicted positive class	True positive (TP)	False positive (FP)
Predicted negative class	False negative (FN)	True negative (TN)

Table 3.3: Confusion matrix table for binary classification [81]

Metric	Formula	Intuitive Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	% of predictions that are correct
Precision	$\frac{TP}{TP+FP}$	% of positive predictions that are correct
Recall / Sensitivity	$\frac{TP}{TP+FN}$	% of actual positives, that are predicted positive
Specificity	$\frac{TN}{TN+FP}$	% of actual negatives, that are predicted negative
F1 Score	$2*\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	Harmonic mean of recall and precision. Single metric that balances their trade-off

Table 3.4: Evaluation metrics, calculated using table 3.3

Then to evaluate the final model on the test set, along with the previously mentioned metrics, ROC curves (comparing specificity and sensitivity) and precision-recall curves are visualised, and corresponding areas under curves are reported (AUC-ROC and AUC-PR) with values near 1 being better. This comprehensive suite of metrics gives a robust assessment of discrimination and generalisation abilities and is the common approach in imbalanced dataset [43]. Additionally, training time is commented on for practicality, ensuring a balance between predictive performance and computational efficiency for implementation in SAS' dynamic test-and-learn research environment.

3.4.3 K-fold validation

To ensure robustness and stability in the evaluation metrics for model selection, a k-fold validation technique [82] was employed. This approach involves partitioning the data into k equal parts. (k-1) partitions were used for training the model and then the performance was validated on the remaining (unseen) partition. This procedure is repeated k times, with a different partition of data serving as the validation set for each iteration. The final evaluation measures (see section 3.4.2) are obtained on the validation set by computing an average of the performance measures over the k iterations. Consequently, all parts of the data have been used for training and validation.

Chapter 4

Results and discussion

4.1 Current State within SAS

4.1.1 Human-classification

The dispatch code (from call handler) forms the human classification baseline. Two observations are made, firstly 56% of drug harm-related cases are initially coded as not-drug harm initially (see recall), potentially indicating difficulty in identifying drug harm due to initial call descriptions, changing conditions, or data noise. Secondly, the significant class imbalance, with only 4.1% cases being related to drug harm, results in misleadingly high accuracy and specificity metrics, and reinforces the importance of investigating the range of metrics such as F1 score, precision and recall.

Accuracy	Precision	Recall	Specificity	F1 Score
96.8%	65.5%	43.6%	99.0%	52.3%

Table 4.1: Performance measures of human classification approach; using dispatch code

4.1.2 Rules based approach

The dataset provided also contains the NFOD flag, described in section 1.2. This flag is currently SAS' best approach to inferring if a patient record is related to drug harm. Since the dataset used does not contain the full set of non drug harm cases, it is only possible to infer the recall. Of the 21k drug harm related calls in 2022, the NFOD flag only successfully identified 22.8% of the cases suggesting much more complexity than identifying a few common causes/treatments for drug harm in the free text.

4.2 Baseline model

As seen in the previous section, there is a big opportunity to use more advanced techniques to improve how SAS identifies cases related to drug harm.

4.2.1 Free-text exploration

Visually investigating the n-gram can help understand patterns within the text data. Figures 4.1, A.6 show commonality amongst the most frequent words, such as 'pt states' and 'pt lying'. Differences are also evident in the bigram and trigrams, such as 'pt taken' and 'taken overdose'. This reinforces the value of TF-IDF transformation method (section 3.2.1) to weigh the common n-grams across the two classes less.

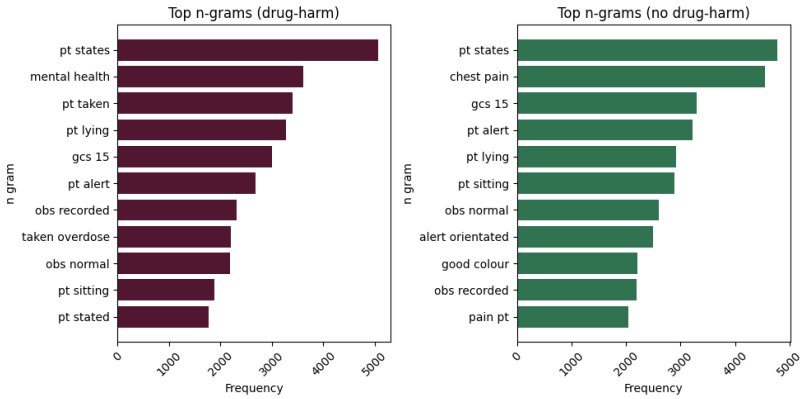


Figure 4.1: Top 10 bigrams, split by drug-harm related or not

Immediately, 'mental health' recurs frequently in drug harm ePRs raising a potential concern when classifying. Where patients have multiple presenting conditions, the code is subjectively given to the primary reason for the emergency. In the SAS system, there is another code for the diagnosis code for mental health (see figure A.1) that can cause noise when trying to learn between only drug-harm and non-drug harm. This will be investigated during evaluation to understand if some of these other presenting factors are causing false positives although there are suggestions of drug-harm.

4.2.2 Logistic regression model

In the sklearn[1] application of logistic regression, the inverse λ value is input, hence the smaller the value, the stronger the regulariser. Through a grid search of values

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

between 0.001 and 100 (base 10 log intervals) and then further grid search between 1 and 10, $\lambda = 4$ returned the highest 10-fold cross-validation F1 score. This λ choice promoted sparsity, reducing the number of features by 99.9% (675k features), down to 466 predictive features for the final model. Results using the balanced test set below indicate a good ability to discriminate between the two classes and therefore interpretation will be expected to be insightful.

Accuracy	**Precision**	**Recall**	**Specificity**	**F1 Score**
91.5%	92.5%	91.5%	91.5%	91.8%

Table 4.2: Evaluation measures of logistic regression on test data, using TF-IDF weighting and unigrams, bigrams and trigrams features, and then L1 regularised with $\lambda = 4$

4.2.3 Interpreting the logistic regression model

As discussed in section 3.2.1, interpreting the coefficients of the logistic regression will help identify what keywords or small phrases are highly predictive of an ePR being related to drug-harm or not, answering RQ2.

The sign of the coefficient indicates if the feature positively or negatively contributes to the probability of the prediction. E.g. using the features list in table 4.3, the presence of words such as ‘overdose’, ‘od’, and ‘take’ increases the probability that the ePR is related to drug harm, and words such as ‘onset’ and ‘ple’ reduces the probability that the ePR is related to drug-harm. The magnitude signifies the importance of the feature, where for example the presence of ’overdose’ in the text increases the probability more than the bigram ’take approx’, whilst both still contributing positively.

Unsurprisingly, words such as ‘overdose’ and ‘od’ are the most important predictive features of the model. The current SAS flag looks at 4 words in the free text: ‘Naloxone’, ‘Narcan’, ‘Heroin’, and ‘Methadone’ which ranked 4th, 5th, 57th and 51st. Verb words such as ‘take’, ‘take approx’ and ‘ingest’ are also highly predictive in the model and hence important additions. Within the negative features, some significant features were interesting to see. For example, ‘seizure’ ranking highly was unexpected since it is a common overdose symptom. Possible reasons include the differences in seizure prevalence between the drug harm and non drug harm classes or also could suggest that if a seizure is seen, whilst this could be due to an overdose, it is instead coded as a code 12 (see figure A.1) - again indicating challenges in the use of single coding for ePRs.

	Positive features		Negative features	
Rank	**Feature**	**Coefficient**	**Feature**	**Coefficient**
1	overdose	80.1	onset	-22.9
2	od	50.4	ple	-19.8
3	take	44.2	pain	-18.5
4	naloxone	38.5	wheeze	-14.0
5	narcan	30.6	seizure	-13.9
6	tablet	26.6	right	-13.5
7	cocaine	25.3	blood	-13.5
8	stimulus	24.5	episode	-12.7
9	bottle	23.8	air	-12.6
10	mg	22.7	day	-11.8
11	take approx	27.6	sob	-11.1

Table 4.3: Extract of feature importance list using coefficients from logistic regression

Moreover, some words in isolation (e.g. ‘tablet’, ‘take’, ’bottle’) do not make intuitive sense for being highly predictive. For example, ‘take’ in isolation is not directly solely linked to drug-harm, this suggests that context words that link ‘take’ to drug harm are likely to be more distant, beyond a trigram feature. This reinforces the advantage of applying transformer models which can capture context of long sentences, due to the self-attention mechanism.

Application

These insights can be used in multiple ways. In situations where high classifier performance is required, then implementing a logistic regression flag, or as seen later, a BERT model, can enhance patient identification to deliver proactive support through, for example, drug-reduction programs. However, when explainability is a necessity, e.g. for governance purposes, an enhanced rule-based NFOD flag is instead required. Positive words from this study can complement the existing set of four words currently used. However, as discussed in section 2.6, these features are correlated with drug harm no causal, hence enhancement decisions must be supplemented with expert guidance from clinicians. For example, as discussed earlier, simply adding the word ‘take’ or creating rules to remove features such as ‘seizure’ and ‘episode’ might not suffice and

lead to an increase in FP and FNs.

4.3 Deep learning classifier model

Sections 4.1 highlighted the opportunity and 4.2.2 demonstrated improvements in drug-harm identification. A baseline logistic regression model demonstrated strong performance with a 10-fold cross validation F1 score of 91.8% on a balanced dataset. Insights into the output, however, suggested that additional performance uplift can obtained by gaining a deeper understanding of the full text files. This section lays out the implementation of BERT through three incremental experiments, initially tuning the hyper-parameters of the language model using only free-text and binary labels to indicate drug-harm, and then testing text-preprocessing techniques to address challenges faced on ePRs captured in an emergency setting.

4.3.1 Experiment 1 - Fine-tuning model with different hyper-parameters

Yinhan et al. [49] showed that hyperparameter selection has a significant impact on the performance of the language models. Therefore, hyperparameter optimisation (HPO) is critical to the build of any language model, and more broadly deep learning models.

The objective for HPO is to minimise the objective function [83];

$$x^* = arg\min_{x \in X} f(x) \tag{4.1}$$

where $f(x)$ is the cross-entropy loss, x^* is the optimal combination of hyperparameters and X is the search space of hyperparameters. Given the fast training times of fine-tuning large language models using GPUs, the general recommended approach to HPO for BERT models is to take a grid search approach (systematically going through each combination of hyperparameters) for hyperparameter selection [16]. The three hyperparameters that are recommended [16] to be fine-tuned along with the suggested ranges are:

Batch size: *16, 32, 64;* ***Learning rate:*** *5e-5, 3e-5, 2e-5;* ***Number of epochs:*** *2, 3, 4*

Keeping batch size at 16 due to memory constraints, this gives 9 combinations of hyperparameters for each of the four models [16, 49, 52, 25], totalling 36 combinations. However, due to computing limitations (NHS laptop with Intel(R) Core(TM) i5-8365U CPU and no GPU acceleration), completing an exhaustive search across the 36 combinations, as suggested by Devlin et.al. [16] is infeasible. For example, fine-tuning

RoBERTa model using batch size 16, learning rate 3e-5 and number of epochs 4 took 38 hours (with 4-fold validation).

Whilst different approaches to HPO on large language models given these constraints have not been explored in previous literature, generally, taking a Bayesian optimisation [84] approach in other machine learning and deep learning tasks have been shown to reach near-optimal solutions in fewer iterations [83] than grid search. Bayesian optimisation [85] learns from past attempts and uses them as additional information when deciding the next combination of hyperparameters to try. Tree-structured Parzen approach (TPE) [86] is a non-standard form of Bayesian optimisation. At a high level, the conditional probability of the hyperparameter combination, given the loss, denoted as $p(x|y)$ is modelled (x is hyperparameter combination, y is loss). Two separate probability densities are constructed using observations, one for 'good' and one for 'bad' (good determined by a threshold). Then the ratio is taken of the 'good' and 'bad' probabilities distributions to determine the hyperparameter combination that has the next highest likelihood of achieving the lowest loss. See Appendix B.2 for a more detailed breakdown and visual representation of how TPE works.

Eggensperger et al. (2013) showed that the Tree-structured Parzen approach performs well on discrete low-dimensional sets where it was able to utilise the recommended hyperparameter choices by experts [84]. Whilst this study's setting involves high dimensional set with BERT models (768 dimensions), inspiration is drawn from the success of using TPE in low dimensional scenarios. The hyperparameter selection choice is guided by using expert-defined options as listed above. For the implementation of this HPO algorithm, Optuna [87] is used, which is an automated optimisation framework. Optuna's framework allows for an simple implementation of HPO, with a range of search strategy choices, of which TPE is the selected search method.

5 combination choices using Optuna were run for each model choice and the performance of best hyperparameter combination is reported below.

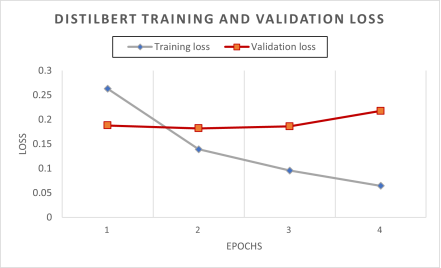
This experiment had expected to see different hyperparameter combinations to be optimal for the different models due to their slightly different architecture, different ways that they were pre-trained, and the pre-training weights. Interestingly, it was observed that the learning rate of 3e-05 and epochs of 2 is the best combination (by cross-entropy loss) for each model.

Learning rate 3e-05 best allows the model to converge on this dataset. Bringing this to life, when using BERT with epoch=2, varying learning rates to 2e-05, 3e-05, and 5e-05 saw F1-scores of 93.0%, 93.7% and 92.8% respectively.

Metric	DistilBERT	BERT	RoBERTa	BioBERT
Loss	0.182	0.174	0.193	0.181
Accuracy	93.1%	93.5%	92.5%	93.2%
Precision	92.4%	91.8%	90.0%	91.8%
Recall	94.0%	95.7%	95.7%	95.2%
Specificity	94.1%	95.2%	95.1%	94.5%
F1 score	93.2%	93.7%	92.8%	93.5%

Table 4.4: Best hyper-parameter combinations were learning rate: 3e-05, epochs: 2 for all models following 5 iterations for each mode, using 4-fold validation.

Epochs Across each model, increasing epochs to greater than 2 saw training loss fall, however validation loss increased - suggesting overfitting



BERT is the best-performing model with lowest loss and higher F1 score. DistilBERT achieved similar performance to BERT (F1 Δ = -0.5%), with higher precision and approximately half the training time (10hrs vs 19hrs). Surprisingly, whilst recall was high, RoBERTa did not perform as well overall. This was an unexpected finding, but could be due to the hyperparameter choices that were tested and that there is no guarantee to find an optimal solution faster than by taking an exhaustive grid search approach, or alternatively the RoBERTa model experienced more overfitting and does not generalise as well to the unseen clinical text-domain. With only hyperparameter tuning, BioBERT did not perform in line with other literature findings where when comparing BERT and BioBERT in [26] BioBERT generalises better on ePRs, suggesting that the additional nuances in this data called out in section 3.1.3 does cause the model to struggle to generalise as well.

4.3.2 Experiment 2 - Domain-specific pre-processing and model

As expected, particularly due to the time-critical nature of emergency callouts, the free text in the study's ePRs sees the use of many acronyms and abbreviations (AA). This experiment explores the impact that providing more context words through the

expansion of AA has on model performance.

The total number of AA within the free-text in the dataset was initially quantified using The Medical Abbreviation and Acronym Meta-Inventory [44] which contains over 100k clinical abbreviations. Conducting a dictionary lookup, 120k words out of a total of 1m words in training were identified to be AA in the free text. Whilst this dataset has several expansions for each AA, as seen in section 2.3.3, automated approaches exist using word vector similarities, but expert knowledge for expansion remains to perform best and is a feasible approach in a relatively small dataset. Therefore, a bespoke dataset was collated for this experiment with one expansion for each AA.

This bespoke dataset was collated using a number of sources [88, 89, 90] and 45 additional AA with >50 instances identified during the initial AA sizing task. Any words that could have a high probability of being a word itself (such as 'so'), or having conflicting expansions were removed. The resulting dataset was formed of 395 AA, with the expansions validated by a clinician. This data is also shared with SAS for other text related tasks. Using this dataset, 94k words (out of 1m) are identified as AA (9%), which are 80% of all AA identified. With the aim to provide BERT with more context words, having an uplift of 9% was deemed suitable and promising at this stage.

Abbreviation	**Count**	**Abbreviation**	**Count**	**Abbreviation**	**Count**
PT	45,944	*GP*	2,492	*PTS*	1,071
T	5,586	*APPROX*	2,491	*SOB*	1,031
O/A	5,576	*ECG*	1,785	*BP*	1,000
O/E	4,527	*Hx*	1,758	*OD*	915
GCS	2,716	*C/O*	1,226	*A&E*	808

Table 4.5: Top 15 acronyms identified in training and test dataset, before anonymisation

The model choices for this experiment were selected to best answer RQ1 addressing multiple angles. BioBERT was tested to see if the full form of specific words such as 'DCA' to 'double crewed ambulance' allows it to learn better. BERT was selected as the best-performing general domain model, and DistilBERT was selected to continue to evaluate performance and training time trade off. The original ePR text was run against the bespoke AA database and then anonymised. 4-fold validation results presented here.

Across all of the models, the precision had increased and significantly for BERT and BioBERT. This suggests that the use of full forms helps the model make fewer incorrect false positive predictions. For BERT and BioBERT, this was offset by a decrease

	DistilBERT		**BERT**		**BioBERT**	
	Base	with A/E	Base	with A/E	Base	with A/E
Loss:	0.182	0.186	0.174	0.185	0.181	0.173
Accuracy:	93.1%	93.2%	93.5%	92.2%	93.2%	93.4%
Precision:	92.4%	92.5%	91.8%	94.7%	91.8%	92.8%
Recall:	94.0%	94.4%	95.7%	89.8%	95.2%	94.4%
Specificity:	94.1%	94.2%	95.2%	89.9%	94.5%	94.0%
F1 score:	93.2%	93.4%	93.7%	92.2%	93.5%	93.6%

Table 4.6: Exp 2 results; using acronym expansion (A/E) compared to Exp 1 (Base)

in recall. While expanding the abbreviations helps to improve the predictions on the minority class, the model is missing more positive cases. This could be because attention is now being paid more on specific words, and therefore is less able to generalise to alternate words that describe the same thing (e.g. seizure and convulsions). Broadly, DistilBERT sees no improvement compared with experiment 1 performance.

Overall, the expansions had limited F1 performance uplift (BioBERT and DistilBERT). Table 4.5 suggests this could be attributed to the AAs identified. By only picking the AA with only one high-probability expansion results in many generic words (>50%) such as PT (patient), O/A (on arrival) and only a few that are distinctly linked to drug harm or not e.g. OD (overdose) and SOB (shortness of breath). This could suggest the dataset approach used was too cautious and hence the expansions are not providing enough distinct information to help predictions. Whilst the word similarity approach is less accurate, it could see overall improvement. Nonetheless, as there were marginal improvements, this additional step was used in the final model build.

4.3.3 Experiment 3 - Adding additional features into the free-text

This experiment tests the hypothesis that providing additional information relevant to the emergency callout, such as time and severity indicators, can enhance the ability of the model to learn context, consequently leading to better predictions (refer to section 3.1 for a list of all features available in dataset). This experiment is particularly important given the lack of specific information such as patient history that is usually used in clinical text classification tasks (see section 3.1.3). To prevent adding unnecessary noise to BERT, and enabling learning from critical context whilst constrained by each ePR

document length of 256 tokens, this experiment starts with feature selection.

Preprocessing of the dataset was required for the features to be able to be used by the models. The categorical features (area, day of the week, and call colour) are converted into a set of binary features using one-hot encoding, and time is broken into hours and minutes. Whilst time is continuous, due to the cyclical nature (e.g. 23hrs is closer to 00hrs than 10hrs), was also one-hot encoded into 8 binary 'hour' features (e.g. 0-3hrs, 3-6hrs) and 6 'minute' features for minutes (e.g. 0-10, 10-20). Post pre-processing, 45 features were available. Note month was removed due to limitation called out in section 3.1.1, but date was used to create 'day of week feature'.

Backward selection (implemented via the sklearn library) selected 18 of the most important features using random forest (RF) models with default settings as the hyper-parameters and with evaluation criterion as the F1 score across 5-folds.

Embedded method The strength of the regulariser λ drives sparsity. Therefore the value of λ was selected to identify the 18 most important features. Through a grid search of values between 0.01 and 1 for λ, the value of 0.022 achieved this. (Figure B.4 shows impact of regulariser on feature numbers)

The aggregation method was the union of the two approaches, resulting in 25 identified features, i.e. if either or both selection approaches identified the feature as important, then it was kept.

Features identified	**Features not identified**
day of week: Fri, Sat, Sun	day of week: Mon, Tue, Wed, Thu
area: EH, PA, KA, ML, AB, FK, G, KY, KW, TD	area: DD, DG, HS, IV, PH, ZE
hour group: 0-18, 21-24	hour group: 18-21
minute group: 10-30	minute group: 0-10, 30-60
call colour: Red, Green, Yellow, Amber, Unknown	call colour: Lime, Purple, No colour

Table 4.7: Final feature selection summary. Area codes expansion in table A.4.

In line with expectations, Friday to Sunday were identified as important days. Several areas were ranked importantly, however, some areas where drug harm is known to be more prevalent such as Dundee were not identified. Here expert judgement from SAS was applied and was therefore included. For hours, 7 out of the 8 features were

identified as important. For ease of implementation, all hours were included. Within call colour, since 'unknown' does not give us any true information on the severity of case for the model to learn from, this feature is manually excluded.

Adding these features caused a shift in the distribution of tokens, with now 16.8% documents exceeding 256 tokens. To avoid any additional information loss, for this experiment, the size of the input vector was increased to 285 tokens, noting the extra training time required, to bring the percentages of cases being truncated back down to 11.0% so that the same context used in experiment 1 and 2 are still retained here (impact of word length on prediction in Appendix B.2).

To conclude, the features from the first column in table 4.7 are included, except Dundee is added and unknown call colour is removed. As an example, "Friday, 23:22, call colour urgency: amber, Location: Edinburgh" appends to the beginning.

Results

With the features appended as free-text to the ePR document and abbreviations expanded, the results are reported below with 4-fold validation.

	DistilBERT			**BERT**			**BioBERT**		
	Base	**A/E**	**A/F**	**Base**	**A/E**	**A/F**	**Base**	**A/E**	**A/F**
Loss	0.182	0.186	0.185	0.174	0.185	0.179	0.181	0.173	0.157
Accuracy	93.1%	93.2%	93.1%	93.5%	92.2%	93.3%	93.2%	93.4%	94.2%
Precision	92.4%	92.5%	90.3%	91.8%	94.7%	93.5%	91.8%	92.8%	93.7%
Recall	94.0%	94.4%	96.4%	95.7%	89.8%	93.2%	95.2%	94.4%	94.2%
Specificity	94.1%	94.2%	89.9%	95.2%	89.9%	93.2%	94.5%	94.0%	93.5%
F1 score	93.2%	93.4%	93.2%	93.7%	92.2%	93.3%	93.5%	93.6%	94.0%

Table 4.8: Comparison of impact of only acronym expansion (A/E) from experiment 2 and A/E with additional features (A/F) for experiment 3 added as free-text on each model.

Compared to Exp 2, the introduction of additional features benefited both BERT and BioBERT to distinguish between the two classes better as seen by an increase in F1 score of 1.1 perc points and 0.4 perc points respectively. Neither experiment saw a big impact in overall performance for DistilBERT. Both experiments have shown benefits in precision for BioBERT with a slightly more cautious model, as seen by a decrease in recall.

4.4 Evaluation of final model

In a practical setting, achieving RQ1 will mean fewer patients who need help (recall) are missed, without driving an increase in ineffective use of resources (precision) by reaching out to the wrong patients. BioBERT was selected for final evaluation due to achieving the highest F1 score of 94.0%.

	Accuracy	**Precision**	**Recall**	**Specificity**	**F1 Score**
NFOD flag	96.4%	64.9%	22.8%	99.5%	33.8%
Logistic regression	89.6%	25.9%	84.2%	89.8%	39.7%
BioBERT fine-tuned	95.1%	44.3%	80.7%	95.7%	57.2%

Table 4.9: Final evaluation on imbalanced test dataset. Benchmark using penalised logistic regression and NFOD flag.

While the model demonstrated strong performance on a balanced dataset (table 4.8), the performance differed significantly when applied to the imbalanced test data. The NFOD flag achieves the highest precision (64.9%) but sees a very low recall (22.8%), meaning many patients are missed. It also had the highest accuracy which was driven due to high specificity (predicting the negative case). Applying logistic regression (LR) sees an improvement in recall (84.2%), however precision is significantly impacted (25.9%), meaning whilst more patients are identified, many non-drug harm patients are also being incorrectly identified which would lead to increased inefficient use of SAS resources. BioBERT, balancing the trade-off between precision and recall well, saw a higher F1 score (+23.4 percentage points higher than NFOD), capturing more patients whilst limiting wasteful outreach. Since the data used has a significant imbalance, analysing ROC and PR curves are important. BioBERT outperforms LR and NFOD flag with a higher ROC AUC (0.967) showing a very strong ability to discriminate between the two classes (a score of 1 signals perfect classifier). Moreover, BioBERT also has the highest AUC-PR value of 0.701, demonstrating a better overall precision-recall balance (vs 0.454 for NFOD flag).

With an improvement in F1 score of +23.4%, BioBERT has shown real improvement in classifier performance than the current rules-based NFOD flag. In real terms, the recall increasing to 80.7% means that if the same performance is applied to full year 2022 data, of the 21.8k drug-harm patients identified through diagnostic code, the NFOD flag identifies 5.0k patients whereas the BioBERT identifies 17.5k patients,

meaning 12.6k fewer patients are missed.

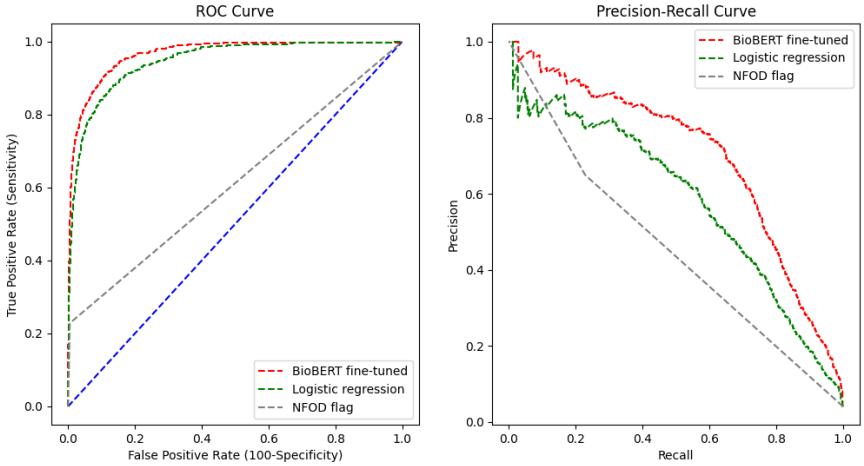


Figure 4.2: ROC curve for BioBERT model, Logistic regression and NFOD flag, with ROC AUC values of 0.967, 0.946 and 0.612 respectively. Precision/Recall curves with AUC-PR values of 0.701, 0.581 and 0.454 respectively.

Investigating table 4.10 to understand performance drop suggests precision and recall fall is primarily driven due to cohorts 2 and 3 (see table 3.2). Whilst the choice to have a balance of cohorts 2 and 4 that was representative of the population was deliberate to avoid bias being brought into the training due to dispatch code (only diagnostic code was used as golden truth), weighting cohorts 2 and 3 more heavily in training as proxies for the more 'complex' could have helped the model understand intricacies better.

Cohort	**TP**	**TN**	**FP**	**FN**
1	83.7%			16.3%
2		61.6%	38.4%	
3	77.7%			22.3%
4		96.0%	4.0%	

Table 4.10: Confusion matrix for each cohort

To understand this further, all of the cohort 2 FP cases were manually reviewed and re-coded, revealing 71.1% should have been labeled as drug-harm related. Examples seen include phrases such as "took an overdose of diazepam" and "patient vomited since taking overdose". The issues arise since the current system only allows paramedics

to assign one code to each ePR, even when multiple presenting conditions exist. For example, of the cohort 2 FP cases, 34.0% were coded as mental health and 26.3% as 'other presenting complaint'. Similarly, inspecting the cohort 4 cases with >90% drug harm probability showed drug harm instances, such as "took unknown amounts of diazepam, sertraline". Therefore, before conclusions are drawn on the true performance of the model, further work is required to correctly label a test sample of cases on if this is drug harm related or not.

Nonetheless, this study builds on the foundations laid by Manca et al. [3] well. This study navigated two additional challenges, namely greater imbalance (4.3% minority class vs 27% in Manca et al. study) which can impact performance outcomes. Moreover, compared to a more narrow definition of alcohol harm, this study faced additional complexity of identification as there are unfortunately many ways by which drug harm can manifest, including illicit drugs, prescriptions, and over-the-counter medication. This is evident by seeing the recall for the NFOD flag at 22.8%, compared to 38.0% for the alcohol flag. Despite the increased complexity, and without any expert manual feature engineering, this study saw a recall uplift of +0.579 in line with their uplift of +0.562 from the alcohol flag up to 0.942. However, it is noted that the specificity for this study dropped by -0.038 where they maintained specificity in line with their alcohol flag. This could be due to differences in class imbalance but further investigation is required.

To summarise, the study's findings demonstrate the potential of using deep learning approaches like BioBERT to significantly enhance capabilities on this type of data. Even with the additional complexities of identifying drug harm in comparison to previous literature on alcohol harm [3] or opioid misuse [29], the model saw real improvements in performance achieving AUC-PR +0.247 higher than the NFOD flag. A key positive finding is that using diagnosis codes to allow the model to learn relationships between the words forms an effective classifier (despite some additional noise due to incorrect labels), and manual inspection has shown that the model has correctly identified several cases of drug-harm in ePRs that were previously not known to SAS. This gives the service another powerful tool to identify the right patients that historically were missed.

Chapter 5

Conclusions

5.1 Main results

This study has demonstrated that there are patterns within the free text that enable the classification of ePRs related to drug harm. Specifically, this project navigated the challenge of no access to in-hospital clinical information to get a holistic view of the patient, which distinguishes it from prior BERT-based studies on in-hospital ePRs. Notably, it also extends past prior work on alcohol harm using ML on a similar dataset within SAS by demonstrating the first application of BERT in this context.

RQ1: Investigating text classifier performance

In practice, the service benefits from improved recall as fewer patients impacted by drug harm are missed, and improved precision as there is lower waste of resources in reaching out to the incorrect patients, therefore the aim was to optimise the F1 score. Two model architectures were implemented to improve on the NFOD flag's performance.

Baseline model: A penalised L1 logistic regression with TF-IDF weighted unigram, bigram and trigrams achieved an F1 score of 39.7%. The recall was at 84.2% however precision performed poorly at 25.9%. Performing well on a balanced set, interpretation of features suggested predictive relationships between words went beyond trigram distances reinforcing the need for a model that is able to learn context from longer sequences. Therefore, the BERT model was investigated.

BERT model choice: Various models were tested to understand their performance on emergency text: general domain models, (BERT and RoBERTa), domain-specific BioBERT, and a more compact DistilBERT model which trains in half the time. Different experiments were used to test the impact of abbreviation expansion and additional

feature use in the absence of other clinical information. Interestingly, BERT generalised best without additional preprocessing, while DistilBERT exhibited stable F1 scores across each experiment. It was found that abbreviation expansion did not see any significant benefit across any of the models, likely because the expanded words were not discriminatory across the classes. Adding additional features did seem to have a strong benefit for BioBERT showing providing additional relevant medical context does see it learn better.

Best model selection: The optimal model was BioBERT with abbreviations expanded and additional features such as time, location and call urgency added as free text. This achieved an F1 score on the unseen dataset of 57.2% (uplift of 23.4% from NFOD flag). Model performance did suffer on imbalanced data (validation F1 score on balanced set saw F1 score of 94.0%). Upon further investigation, data quality was noticed to be an issue in cases where there could be multiple presenting conditions (such as overdose and mental health). For a robust evaluation of true performance, further work is required to move from the current diagnostic code as golden source to a manually created dataset. Nonetheless, this classifier has learned effectively from the diagnostic code labelling (with additional noise where labels are incorrect) which has also identified ePRs related to drug harm that were previously not known to SAS.

RQ2: Interpreting keywords that are predictive of drug harm cases

The use of the penalised L1 logistic regression model was so direct context-free words are phrases can be identified to directly improve the current NFOD flag, ensuring explainability given governance requirements in reporting figures externally. Using the penalised L1 logistic regression, the number of features used were reduced by 99.9% down to 466 features which achieved a F1 score of 91.8% on a balanced test set, showing ability to discriminate between the classes. These features consisted of unigrams, bigrams and trigrams. The four words currently used in the NFOD flags are ”Naloxone”, ”Narcan”, ”Heroin”, and ”Methadone”, which rank 4th, 5th, 57th and 51st. This suggests that there are a greater set of words and phrases that can be included to enhance the NFOD flag and improve recall. Simple implementations could include words such as od, overdose, cocaine. Some words may be challenging to introduce in a rules based approach, given the dependency on other words, such as ’take’, ’tablet’, however for these words bigrams and trigrams could be more useful such as ’take approx’, ’over 40 tablets’. Given the co-dependencies of the words, an expert overlay is required to ensure noise is not being added into the flag.

Limitations

The two main limitations of this study were due to hardware constraints and data quality. Without access to GPUs, training times were significantly slower, limiting the ability to fully search through all hyperparameter to find the optimal combinations. This constraint led to the use of TPE which has shown promise in literature, however could have impacted the poorer performance of RoBERTa. The second limitation was related to the coding of the golden truth, i.e. the diagnostic code. In the absence of being able to add multiple codes to an ePR, paramedics are required to assign a code depending on what the most presenting condition is, bringing inconsistencies with how the ePRs are coded. A refinement of the golden standard coding is essential to gain a better understanding of the classification model's true performance.

5.2 Future work

This section outlines the key impactful avenues that can be extended from this research.

Investigating multi-label classification The current system assumes mutual exclusivity among conditions due to the single diagnostic code constraint per case. Investigations into false positives revealed cases with multiple co-occurring conditions (e.g. mental health and overdose), suggesting this assumption might be too strong. Exploring multi-label classification, where each ePR has multiple codes attached, could uncover a much more comprehensive understanding of patients' conditions, surfacing conditions that may be currently understated.

Hyperparameter optimisation BERT model's performance is heavily impacted by hyperparameter choice. Exploring the impact of optimisation techniques on language models could guide similar hardware-constrained studies involving large language models. An interesting area to explore would be to understand the number of iterations usually required for TPE algorithm (used in experiment 1) to reach near-optimal performance compared to an exhaustive grid search.

BERT interpretability This study investigated interpretability on a context free model for purposes of guiding direct improvements into SAS' rules-based NFOD flag. As BioBERT enhanced performance, exploring interpretability could offer qualitative insights into additional contextual nuances to what the patient is facing at times of emergency that can help the service tailor responses.

Bibliography

[1] Hannah Ritchie, Pablo Arriagada, and Max Roser. Opioids, cocaine, cannabis and other illicit drugs. *Our World in Data*, 2022.

[2] Institute for Health Metrics and Evaluation, Global Burden of Disease (2019). Available: https://www.healthdata.org/research-analysis/gbd [Accessed: 07-08-2023].

[3] Francesco Manca, Jim Lewsey, Ryan Waterson, Sarah M. Kernaghan, David Fitzpatrick, Daniel Mackay, Colin Angus, and Niamh Fitzgerald. Estimating the Burden of Alcohol on Ambulance Callouts through Development and Validation of an Algorithm Using Electronic Patient Records. *International Journal of Environmental Research and Public Health*, 18(12):6363, 6 2021.

[4] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A Survey on Text Classification: From Shallow to Deep Learning. 8 2020.

[5] Yaakov HaCohen-Kerner, Daniel Miller, and Yair Yigal. The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*, 15(5), 5 2020.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. 1 2013.

[7] Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, pages 29–48, 2003.

[8] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, 752:41–48, 1998.

[9] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. pages 137–142. 1998.

[10] Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, 5(1):12, 12 2020.

[11] Alexander Genkin, David D Lewis, and David Madigan. Sparse logistic regression for text categorization. *DIMACS Working Group on Monitoring Message Streams Project Report*, 2005.

[12] Nasir Jalal, Arif Mehmood, Gyu Sang Choi, and Imran Ashraf. A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University - Computer and Information Sciences*, 34(6):2733–2742, 6 2022.

[13] Berna Altınel and Murat Can Ganiz. Semantic text classification: A survey of past and recent advances. *Information Processing & Management*, 54(6):1129–1153, 11 2018.

[14] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep Learning–based Text Classification. *ACM Computing Surveys*, 54(3):1–40, 4 2022.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. 6 2017.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10 2018.

[17] Olaronke G. Iroju and Janet O. Olaleke. A Systematic Review of Natural Language Processing in Healthcare. *International Journal of Information Technology and Computer Science*, 7(8):44–50, 7 2015.

[18] Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, 116:494–520, 2 2019.

[19] Irena Spasic and Goran Nenadic. Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics*, 8(3):e17984, 3 2020.

[20] Irena Spasić, Jacqueline Livsey, John A. Keane, and Goran Nenadić. Text mining of cancer-related information: Review of current status and future directions. *International Journal of Medical Informatics*, 83(9):605–623, 9 2014.

[21] H. Yang, I. Spasic, J. A. Keane, and G. Nenadic. A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries. *Journal of the American Medical Informatics Association*, 16(4):596–600, 7 2009.

[22] Jonathan Bates, Samah J Fodeh, Cynthia A Brandt, and Julie A Womack. Classification of radiology reports for falls in an HIV study cohort. *Journal of the American Medical Informatics Association*, 23(e1):e113–e117, 4 2016.

[23] Tak Sung Heo, Yu Seop Kim, Jeong Myeong Choi, Yeong Seok Jeong, Soo Young Seo, Jun Ho Lee, Jin Pyeong Jeon, and Chulho Kim. Prediction of Stroke Outcome Using Natural Language Processing-Based Machine Learning of Radiology Report of Brain MRI. *Journal of Personalized Medicine*, 10(4):286, 12 2020.

[24] Hoang Nguyen and Jon Patrick. Text Mining in Clinical Domain: Dealing with Noise. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 549–558, New York, NY, USA, 8 2016. ACM.

[25] 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, 1 2019.

[26] Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. Comparative Analysis of Text Classification Approaches in Electronic Health Records. 5 2020.

[27] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly Available Clinical BERT Embeddings. 4 2019.

[28] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *BioNLP 2019 - SIGBioMed Workshop on Biomedical Natural Language Processing, Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, 6 2019.

[29] José Tomás Prieto, Kenneth Scott, Dean McEwen, Laura J Podewils, Alia Al-Tayyib, James Robinson, David Edwards, Seth Foldy, Judith C Shlay, and Arthur J Davidson. The Detection of Opioid Misuse and Heroin Use From Paramedic Response Documentation: Machine Learning for Improved Surveillance. *Journal of Medical Internet Research*, 22(1):e15645, 1 2020.

[30] Svetlana Kiritchenko and Saif M. Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. 5 2018.

[31] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. 6 2019.

[32] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

[33] Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344, 2008.

[34] Aitor García-Pablos, Naiara Perez, and Montse Cuadros. Sensitive Data Detection and Classification in Spanish Clinical Text: Experiments with BERT. 3 2020.

[35] Sravana Reddy and Kevin Knight. Obfuscating Gender in Social Media Writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26, Stroudsburg, PA, USA, 2016. Association for Computational Linguistics.

[36] Erenay Dayanik and Sebastian Padó. Masking Actor Information Leads to Fairer Political Claims Detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[37] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 6 2004.

[38] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 4 2009.

[39] Tuanfei Zhu, Yaping Lin, and Yonghe Liu. Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recognition*, 72:327–340, 12 2017.

[40] Zubair Afzal, Martijn J Schuemie, Jan C van Blijderveen, Elif F Sen, Miriam CJM Sturkenboom, and Jan A Kors. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC Medical Informatics and Decision Making*, 13(1):30, 12 2013.

[41] Ryan Sullivan, Robert Yao, Randa Jarrar, Jeffrey Buchhalter, and Graciela Gonzalez. Text Classification towards Detecting Misdiagnosis of an Epilepsy Syndrome in a Pediatric Population. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2014:1082–7, 2014.

[42] V. García, J.S. Sánchez, and R.A. Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2 2012.

[43] Hongxia Lu, Louis Ehwerhemuepha, and Cyril Rakovski. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology*, 22(1):181, 12 2022.

[44] Lisa Grossman Liu, Raymond H. Grossman, Elliot G. Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K. Vawdrey. A deep database of

medical abbreviations and acronyms for natural language processing. *Scientific Data*, 8(1):149, 6 2021.

[45] Hua Xu, Peter D Stetson, and Carol Friedman. A study of abbreviations in clinical notes. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2007:821–5, 10 2007.

[46] Sergeui Pakhomov, Ted Pedersen, and Christopher G Chute. Abbreviation and acronym disambiguation in clinical discourse. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2005:589–93, 2005.

[47] Yue Liu, Tao Ge, Kusum S. Mathews, Heng Ji, and Deborah L. McGuinness. Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion. 4 2018.

[48] Lisa Torrey and Jude Shavlik. Transfer Learning. In *Handbook of Research on Machine Learning Applications and Trends*, pages 242–264. IGI Global, 2010.

[49] 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. 7 2019.

[50] Alexander Turchin, Stanislav Masharsky, and Marinka Zitnik. Comparison of BERT implementations for natural language processing of narrative medical documents. *Informatics in Medicine Unlocked*, 36:101139, 2023.

[51] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, New York, NY, USA, 8 2006. ACM.

[52] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 10 2019.

[53] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. pages 1631–1642.

[54] Warren J. von Eschenbach. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34(4):1607–1622, 12 2021.

[55] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[56] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch. 9 2020.

[57] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 5 2019.

[58] Cynthia Rudin. Why black box machine learning should be avoided for high-stakes decisions, in brief. *Nature Reviews Methods Primers*, 2(1):81, 10 2022.

[59] Guohou Shan, James Foulds, and Shimei Pan. Causal Feature Selection with Dimension Reduction for Interpretable Text Classification. 10 2020.

[60] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at {HLT}-{NAACL} 2003*, pages 142–147. 2003.

[61] Sue M. Evans, Angela Murray, Ian Patrick, Mark Fitzgerald, Sue Smith, Nick Andrianopoulos, and Peter Cameron. Assessing clinical handover between paramedics and the trauma team. *Injury*, 41(5):460–464, 5 2010.

[62] Xinyu Dong, Sina Rashidian, Yu Wang, Janos Hajagos, Xia Zhao, Richard N Rosenthal, Jun Kong, Mary Saltz, Joel Saltz, and Fusheng Wang. Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2019:389–398, 2019.

[63] Vimala Balakrishnan and Lloyd-Yemoh Ethel. Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, 2(3):262–267, 2014.

[64] Alexander M. Robertson and Peter Willett. Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1):48–67, 3 1998.

[65] William B Cavnar and John M Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Las Vegas, NV, 1994.

[66] Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. The use of bigrams to enhance text categorization. *Information Processing & Management*, 38(4):529–546, 7 2002.

[67] Aaron J. Masino, Robert W. Grundmeier, Jeffrey W. Pennington, John A. Germiller, and E. Bryan Crenshaw. Temporal bone radiology report classification using open source machine learning and natural langue processing libraries. *BMC Medical Informatics and Decision Making*, 16(1):65, 12 2016.

[68] Ben J Marafino, Jason M Davies, Naomi S Bardach, Mitzi L Dean, and R Adams Dudley. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *Journal of the American Medical Informatics Association*, 21(5):871–875, 9 2014.

[69] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1 1996.

[70] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, pages 1157–1182, 3 2003.

[71] R Muthukrishnan and R Rohini. LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pages 18–20. IEEE, 10 2016.

[72] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[73] Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. 3 2019.

[74] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. 9 2016.

[75] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. 11 2017.

[76] Donghai Guan, Weiwei Yuan, Young-Koo Lee, Kamran Najeebullah, and Mostofa Kamal Rasel. A Review of Ensemble Learning Based Feature Selection. *IETE Technical Review*, 31(3):190–198, 5 2014.

[77] Naoual El Aboudi and Laila Benhlima. Review on wrapper feature selection approaches. In *2016 International Conference on Engineering & MIS (ICEMIS)*, pages 1–5. IEEE, 9 2016.

[78] Patrick Juola. Cross-entropy and linguistic typology. In *New Methods in Language Processing and Computational Natural Language Learning*. 1998.

[79] Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. Small and Practical BERT Models for Sequence Labeling. 8 2019.

[80] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. DocBERT: BERT for Document Classification. 4 2019.

[81] Hossin M and Sulaiman M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):01–11, 3 2015.

[82] Tzu Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 9 2015.

[83] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 11 2020.

[84] Katharina Eggensperger, Matthias Feurer, Frank Hutter, James Bergstra, Jasper Snoek, Holger H Hoos, and Kevin Leyton-Brown. Towards an Empirical Foundation for Assessing Bayesian Optimization of Hyperparameters.

[85] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. 12 2010.

[86] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs K{\'e}gl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

[87] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, New York, NY, USA, 7 2019. ACM.

[88] Jennifer Betts. Basic EMS Medical Abbreviations and Acronyms. Available: https://www.yourdictionary.com/articles/ems-acronyms [Accessed: 20-07-2023], 3 2020.

[89] North West Ambulance Service NHS Trust. NWAS Acronym Log. Available: https://www.nwas.nhs.uk/publications/acronym-log/ [Accessed: 21-Jul-2023], 9 2021.

[90] OpenAI. ChatGPT. Question asked: [”Can you provide a list of common abbreviations that paramedics could use in writing their emergency reports”] From: https://chat.openai.com/ [Accessed: 26-07-2023].

[91] Shuhei Watanabe. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. 4 2023.

[92] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 1 2017.

Appendix A

SAS Dataset Explained

A.1 Definitions

MPDS Diagnostic Code
01 Abdominal pain
02 Allergies
03 Animal bites
04 Assault
05 Back pain
06 Breathing problems
07 Burns
08 CO / Inhalation
09 Cardiac arrest
10 Chest pain
11 Choking
12 Seizures
13 Diabetic
14 Drowning
15 Electrocution
16 Eye injury
17 Falls
18 Headache
19 Heart problems
20 Heat / cold exposure
21 Haemorrhage
22 Inaccessible incident
23 Overdose
24 Pregnancy
25 Mental health
26 Generally unwell
27 Penetrating trauma
28 Stroke
29 Road traffic collision
30 Traumatic injury
31 Collapse / unconscious
32 Unknown
33 Interhospital transfer / palliative
45 Healthcare provider emergency call
99 Other presenting complaint

Figure A.1: Scottish Ambulance Service Diagnostic codes

Field	Description
	Integer between 1 and 4, describing dispatch and diagnostic code.
Cohort	Cohort 1 - dispatch OD, diagnostic OD, Cohort 2 - dispatch OD, diagnostic NoOD,
	Cohort 3 - dispatch No OD, diagnostic OD, Cohort 4 - dispatch No OD, diagnostic No OD
Call number	unique identifier of call
Date	The date the call started
Time	The time the call started
Call despatch code	Code assigned by call-handler at point of receiving emergency call. Full list of codes in figure A.1
Diagnostic code	Code assigned by paramedic who attends to patient at the scene. Full list of codes in figure A
Call colour	Urgency of call (which dictates response times). In order; green (least severe), yellow, amber, red, purple (most severe)
	This flag identifies drug harm if the paramedic has checked a box signalling: 'naloxone was given',
NFOD flag	or the 'substance affecting condition' is 'opioids' or 'street benzodiazepine', or if any of the four
	following words are present in the free-text; 'Naloxone', 'Methadone', 'Narcan' and 'Heroin'
Naloxone mentioned	This flag identifies if the words 'naloxone' or 'narcan' are identified in the free text
Heroin mentioned	This flag identifies if the words 'heroin' or 'methadone' are identified in the free text
Additional comments	The free text captured by paramedics when they attend patient at the scene
Postcode	The postcode for ambulance callout
Receiving Hospital	The name of the hospital that the patient is taken to

Table A.1: Description of all fields in SAS dataset

A.2 Deep dive into Cohort 2 and 3

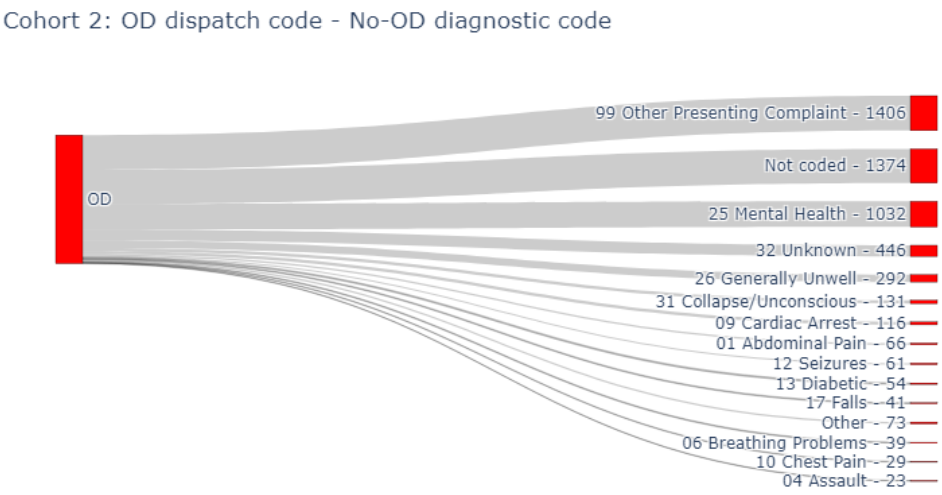


Figure A.2: Sankey diagram of 12.3k cohort 3 cases representing split of initial diagnosis that were then categorised as overdose in diagnostic code

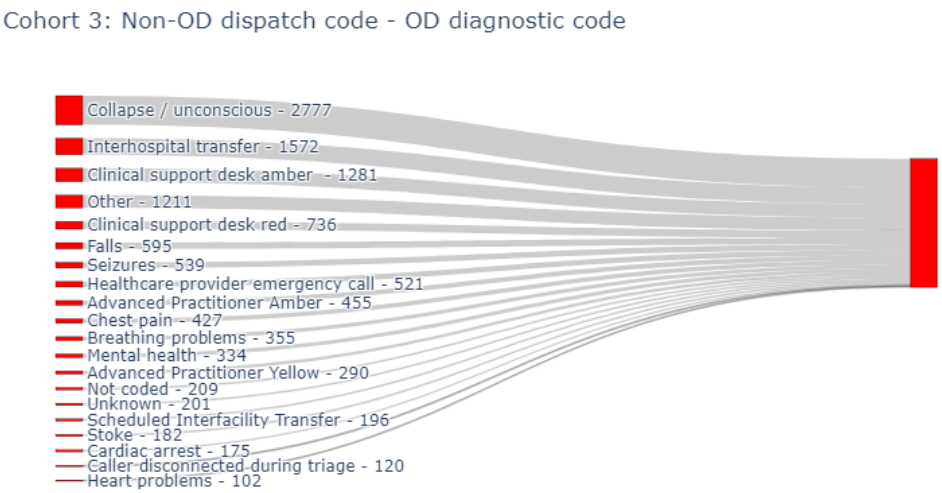


Figure A.3: Sankey diagram of Cohort 2 representing split of initially diagnosed overdoses that were then re-categorised to another diagnostic code

A.3 Free text characteristics

	Drug harm ePR	No drug harm ePR	Delta
Mean number of words	108	117	-9
Median number of words	99	108	-9
Mean text character length	618	668	-50
Median text character length	566	622	-56
Mean number of unique words	82	88	-6
Median number of unique words	79	85	-6
Mean number of duplicated words	25	29	-3
Median number of duplicated words	19	23	-4

Table A.2: Descriptive characteristics of free text in SAS ePR

Tables and graph below show drug harm ePRs have slightly lower number of word, unique words and duplicated words, with spread of data also lower. Given the size of the delta, this should not be an issue.

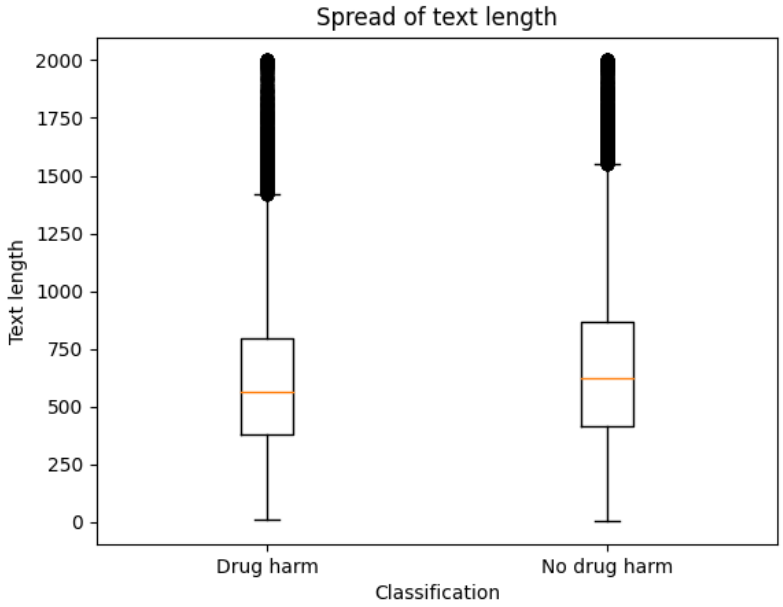


Figure A.4: Spread of ePR character length by ePR. No major difference, no drug harm has larger range.

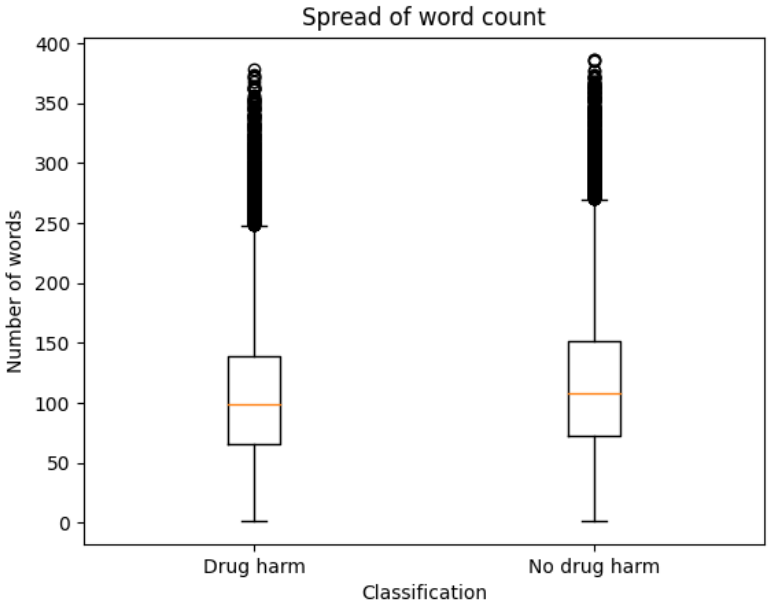


Figure A.5: Spread of ePR word count length by ePR. No major difference, no drug harm has larger range.

A.4 Trigram frequent counts

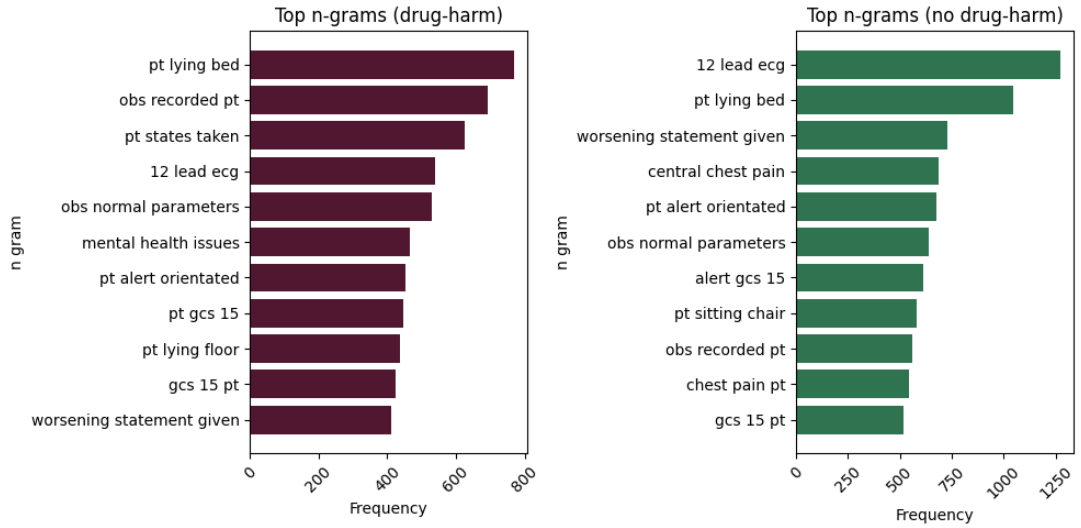


Figure A.6: Top 10 Trigrams, split by drug-harm related or not

A.5 NER categories[1]

Abbreviation	Description
O	Outside of a named entity
B-MIS	Beginning of a miscellaneous entity
I-MIS	Right after another miscellaneous entity
B-PER	Beginning of a person's name
I-PER	Right after another person's name
B-ORG	Beginning of an organization
I-ORG	Right after another organization
B-LOC	Beginning of a location
I-LOC	Right after another location

Table A.3: Full list of NER categories[1]

[1] https://huggingface.co/dslim/bert-base-NER

A.6 Area codes

Postcode area	Area covered
AB	Aberdeen
DD	Dundee
DG	Dumfries and Galloway
EH	Edinburgh
FK	Falkirk and Stirling
G	Glasgow
HS	Outer Hebrides
IV	Inverness
KA	Kilmarnock
KW	Kirkwall
KY	Kirkcaldy
ML	Motherwell
PA	Paisley
PH	Perth
TD	Galashiels
ZE	Lerwick

Table A.4: Scotland areas, by area code

Appendix B

Supplementary information on experiments

B.1 BERT implementation example

B.1.1 The BERT architecture for classification

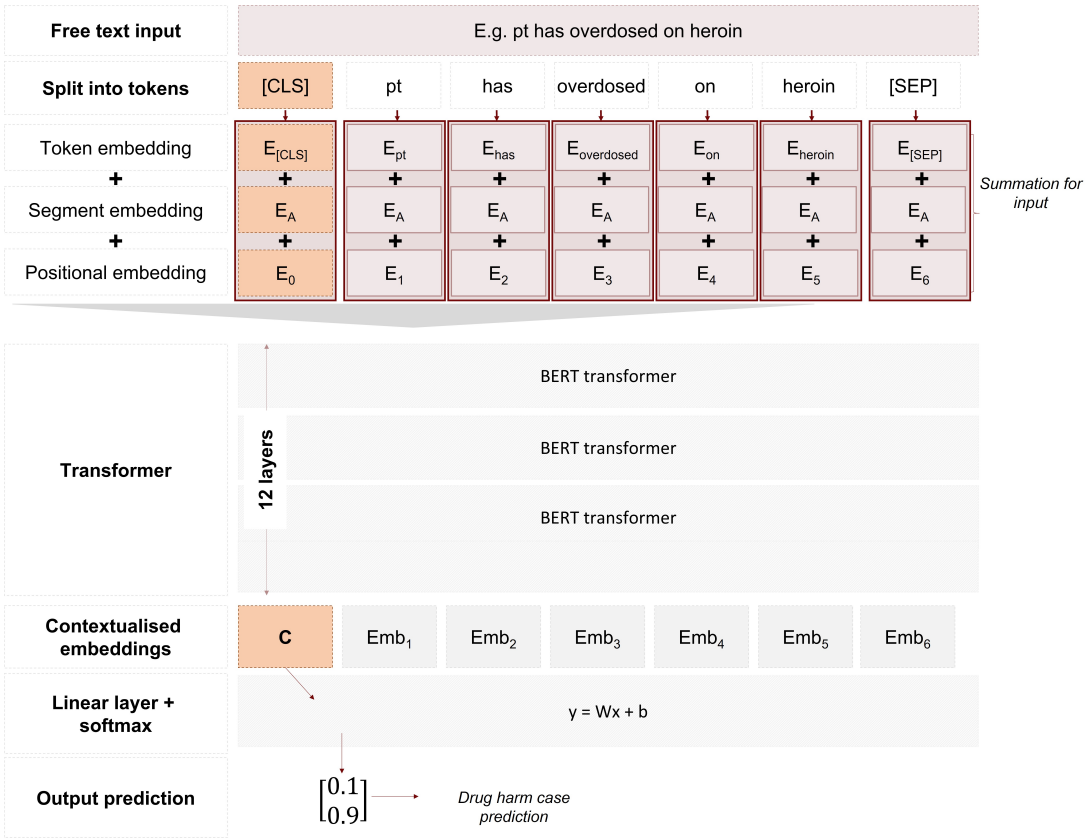


Figure B.1: BERT architecture, adapted from [16]

B.1.2 Creating the input layer

For example, ”Patient given 400mcg naloxone. Confirmed use of opioids” will be processed as follows:

Tokenized:

[CLS], ’patient’, ’given’, ’400’, ’##m’, ’##c’, ’##g’, ’na’, ’##lo’, ’##xon’, ’##e’, ’.’, ’confirmed’, ’use’, ’of’, ’op’, ’##io’, ’##ids’, [SEP]

Token embedding using WordPiece vocabulary:

[101, 5776, 2445, 4278, 12458, 2290, 6583, 4135, 22500, 2063, 1012, 4484, 2224, 1997, 6728, 3695, 9821, 102, 0, 0, ..., 0]

Attention mask:

[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ..., 0]

B.2 Experiment 1 - Tree-parzen Structured estimation

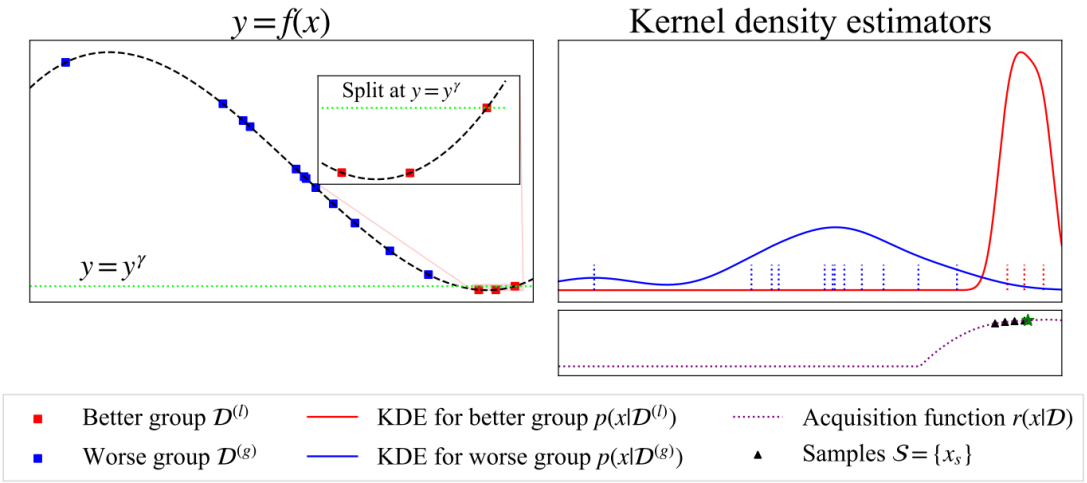


Figure B.2: Visualisation of Tree-structure Parzen Estimator by Watanabe [91]. **Left:** Green dashed line is the threshold between good and bad results. Black dashed line is the cross-entropy loss function, blue and red dots are observations. **Right top:** kernel density estimators using blue and red observations. **Right bottom:** acquisition function, taking a ratio of the good group and bad group (from right top), to help determine the next most promising choice (green star).

TPE takes the following steps:

1. $p(x|y)$ is modelled, where y is the cross-entropy loss and x is the selected hyperparameter combination. Intuitively, this is used to predict the hyperparameter combination, given a loss value.

2. Two probability densities are then created using kernel density estimators [92], where $p(x|D^{(l)})$ uses our 'good' observations (i.e. loss values less than the threshold y^{γ}) and $p(x|D^{(g)})$ uses our 'bad' observations (i.e. loss values that are higher/worse than y^{γ})

$$p(x|y,D) = \begin{cases} p(x|D^{(l)}) & \text{if } y < y^{\gamma} \\ p(x|D^{(g)}) & \text{if } y \geq y^{\gamma} \end{cases} \tag{B.1}$$

3. The acquisition function then uses $\frac{p(x|D^{(l)})}{p(x|D^{(g)})}$ to determine which combination of hyperparameters is most promising to choose for the next iteration. This is because a good 'x' will have a high $p(x|D^{(l)})$ and a low $p(x|D^{(g)})$, making the fraction larger.

B.3 Experiment 2 - Abbreviations

Abbreviation	Count	Abbreviation	Count	Abbreviation	Count
PT	45944	*ED*	643	*mg*	199
T	5586	*NSR*	633	*MH*	194
O/A	5576	*OE*	623	*HxPC*	188
O/E	4527	*LOC*	535	*PRF*	186
GCS	2716	*HR*	489	*ST*	174
GP	2492	*SpO2*	446	*NEB*	156
APPROX	2491	*HPC*	423	*NPA*	135
ECG	1785	*IM*	382	*MI*	132
Hx	1758	*o/d*	373	*min*	131
C/O	1226	*PRU*	355	*CA*	120
PTS	1071	*M*	349	*TIA*	116
SOB	1031	*PMHx*	330	*CVA*	110
BP	1000	*F*	323	*AE*	109
OD	915	*PMH*	317	*PE*	101
A&E	808	*COPD*	315	*C-SPINE*	100
YOM	740	*CPR*	299	*NKDA*	*95*
OA	738	*Tx*	266	*PC*	83
YOF	732	*AF*	266	*SORT*	75
RR	680	*g*	240	*MED*	74
IV	648	*YO*	215	*kg*	71

Table B.1: Top 60 acronyms identified in training and test dataset, before anonymisation

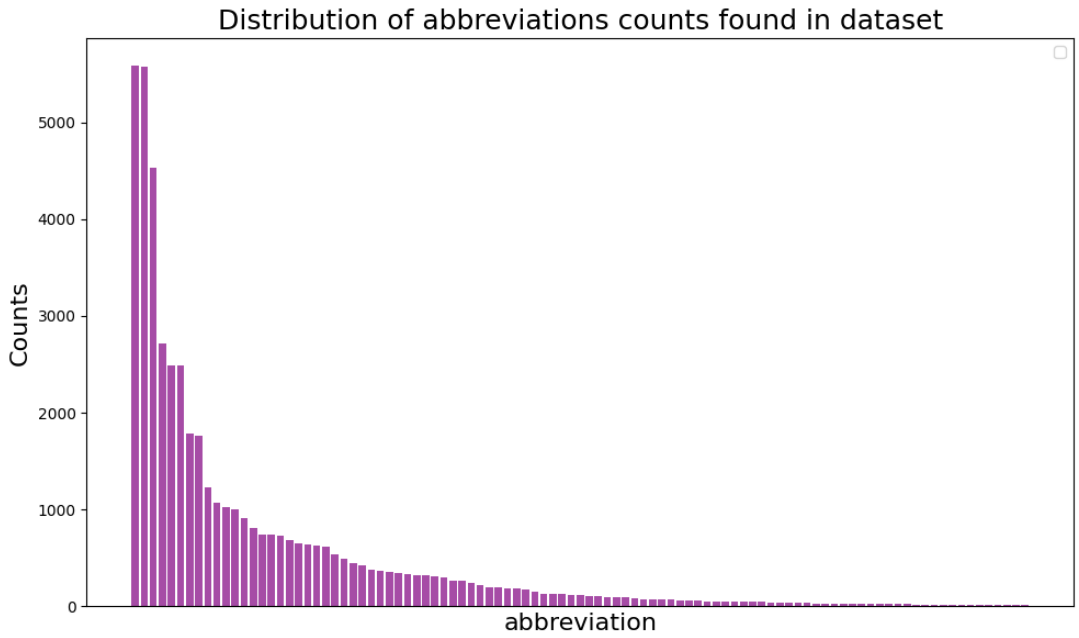


Figure B.3: Frequency count of abbreviations - visually looks like it is following Zipf law

B.4 Experiment 3 - Additional features

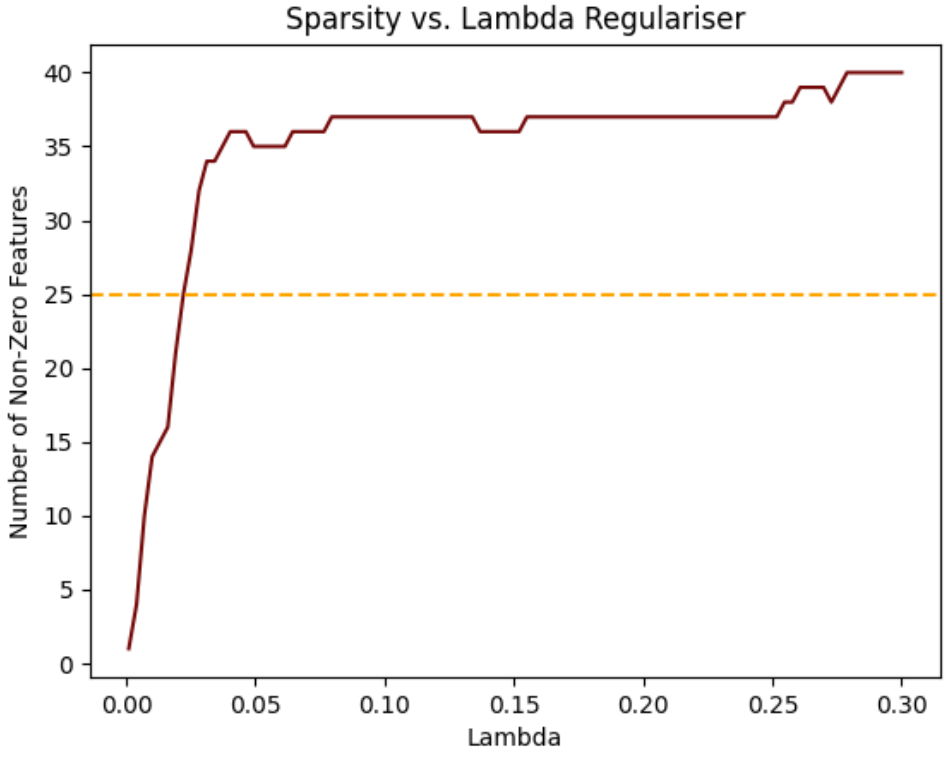


Figure B.4: Impact of regulariser on sparsity. Can see that it is highly sensitive, hence careful tuning is required.

B.5 Evaluation of final performance

Performance split by ePR length

ePR length (words)	FN	FP	TN	TP	Number of cases
1-50	0.80%	11.90%	83.80%	3.50%	1200
51-100	0.30%	5.00%	89.70%	4.90%	3887
101-150	0.30%	3.80%	92.20%	3.70%	4738
151-200	0.10%	3.50%	93.40%	3.00%	2865
201-250	0.20%	3.70%	93.30%	2.80%	1276
251-300	0.20%	3.60%	93.60%	2.70%	450
301-350		0.60%	94.80%	4.60%	174
351-400		6.90%	86.10%	6.90%	72
401-450			100.00%		4

Table B.2: Confusion matrix split by ePR length. There does not seem to be a material impact on performance for ePRs that have been truncated. There, however, is worse performance for shorter length ePRs.