Mining vaccine adverse events mentions from social media using Twitter as a source

Sedigheh Khademi Habibabadi, Pari Delir Haghighi, Frada Burstein, Jim Buttery

Submitted to: JMIR Medical Informatics
on: October 16, 2021

Table of Contents

Mining vaccine adverse events mentions from social media using Twitter as a source

Sedigheh Khademi Habibabadi[1, 2*] PhD; Pari Delir Haghighi[2*] PhD; Frada Burstein[2] Prof Dr; Jim Buttery[1*] Prof Dr

[1]Murdoch Children's Research Institute Melbourne AU
[2]Monash University Melbourne AU
[*]these authors contributed equally

Corresponding Author:
Sedigheh Khademi Habibabadi PhD
Murdoch Children's Research Institute
50 Flemington Rd, Parkville VIC 3052
Melbourne
AU

Abstract

Background: Traditional monitoring for Adverse Events Following Immunisation (AEFI) relies on various established reporting systems, where there is inevitably a lag between an AEFI occurring and its potential reporting, and subsequent processing of reports. AEFI safety signal detection strives to detect AEFI as early as possible, ideally close to real-time. Monitoring social media data holds promise as a resource for this.

Objective: 1) To investigate the utility of monitoring social media for gaining early insights into vaccine safety issues, by extracting vaccine adverse event mentions (VAEM) from Twitter using natural language processing (NLP) techniques. 2) To document the NLP processes used and identify the most effective of them for successively identifying tweets that contain VAEM, with a view to defining an approach that might be applicable to other similar social media surveillance tasks.

Methods: A VAEM-Mine method was developed that combines topic modelling with classification techniques to extract maximal VAEM posts from a vaccine-related Twitter stream, with a high degree of confidence. The approach does not require a targeted search for specific vaccine reactions, but instead identifies any VAEM post within many unrelated posts.

Results: The VAEM-Mine method successively isolates vaccine adverse event mentions from the massive amount of other vaccine-related Twitter posts, achieving an F1-Score of 0.91 in the classification phase.

Conclusions: Social media can assist with detection of vaccine safety signals as a valuable complementary source for monitoring mentions of vaccine adverse events. A social media based VAEM data stream can be assessed for changes to detect possible emerging vaccine safety signals, helping to address the well-recognised limitations of passive reporting systems, including timeliness and under-reporting.

(JMIR Preprints 16/10/2021:34305)
DOI: https://doi.org/10.2196/preprints.34305

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

1. Original Paper

Authors:
Sedigheh Khademi Habibabadi
Faculty of Information Technology
Monash University
Victoria, Australia

Murdoch Children's Research Institute
Melbourne, Victoria, Australia

Email: Sedigh.khademi@mcri.edu.au

Pari Delir Haghighi
Faculty of Information Technology
Monash University
Victoria, Australia
Email: Pari.Delir.Haghighi@monash.edu

Frada Burstein
Faculty of Information Technology
Monash University
Victoria, Australia
Email: Frada.Burstein@monash.edu

Jim Buttery
Professor of Child Health Informatics
Paediatrics Royal Children's Hospital
Victoria, Australia

Murdoch Children's Research Institute
Melbourne, Victoria, Australia

Email: jim.buttery@unimelb.edu.au

Corresponding author:
Sedigheh Khademi Habibabadi

Title: Mining vaccine adverse events mentions from social media using Twitter as a source

Abstract

Background: Traditional monitoring for Adverse Events Following Immunisation (AEFI) relies on various established reporting systems, where there is inevitably a lag between an AEFI occurring and its potential reporting, and subsequent processing of reports. AEFI safety signal detection strives to detect AEFI as early as possible, ideally close to real-time. Monitoring social media data holds promise as a resource for this.
Objectives: 1) To investigate the utility of monitoring social media for gaining early insights into vaccine safety issues, by extracting vaccine adverse event mentions (VAEM) from Twitter using natural language processing (NLP) techniques. 2) To document the NLP processes used and identify the most effective of them for successively identifying tweets that contain VAEM, with a view to defining an approach that might be applicable to other similar social media surveillance tasks.
Materials and Methods: A VAEM-Mine method was developed that combines topic modelling with classification techniques to extract maximal VAEM posts from a vaccine-related Twitter stream, with a high degree of confidence. The approach does not require a targeted search for specific vaccine reactions, but instead identifies any VAEM post within many unrelated posts.
Results: The VAEM-Mine method successively isolates vaccine adverse event mentions from the massive amount of other vaccine-related Twitter posts, achieving an F1-Score of 0.91 in the classification phase.
Conclusion: Social media can assist with detection of vaccine safety signals as a valuable complementary source for monitoring mentions of vaccine adverse events. A social media based VAEM data stream can be assessed for changes to detect possible emerging vaccine safety signals, helping to address the well-recognised limitations of passive reporting systems, including timeliness and under-reporting.

Key words: vaccine reactions; vaccine safety surveillance; social media; Twitter; Machine learning

Introduction

High levels of vaccination uptake are required to effectively immunize a population. Vaccine safety is a key component of effective vaccine delivery, and the ongoing confidence needed for continued high levels of vaccine uptake [1]. Currently, the global importance of vaccines is highlighted more than ever, as the world combats the COVID-19 pandemic. The safety of vaccines is vital, both for vaccine providers and recipients and for public trust and confidence in vaccine programs [2]. Vaccine safety relies upon rigorous compliance to development and manufacturing standards, well conducted clinical trials, thorough assessment, licencing, control, and administration of vaccines. Post-licensure monitoring for vaccine reactions is a key component of ensuring vaccine safety [3].

Vaccine safety surveillance continues in a variety of forms after regulatory approval. It is the primary mechanism to identify serious (and rare) adverse events following immunization (AEFI) that are unlikely to have been exposed by pre-licensure trials, and also allows surveillance in populations that were unable to be included in the trials [4]. Passive (spontaneous) surveillance systems typically rely on spontaneous reporting of adverse events following immunization (AEFI) by individuals and/or by their treating health professionals. They are the main method for gathering Adverse Drug Reactions – they have proven useful in early detection of vaccine and drug related safety issues [5,6]. Although these systems are the backbone of drug safety monitoring, they suffer from major disadvantages, including underreporting, incomplete data, and time lag between an event happening and subsequent reporting of it [7] . The US Vaccine Adverse Events Reporting System (VAERS) is a spontaneous reporting system that anyone can report to, including vaccinees and their families. Health professionals are obligated to report only certain AEFI to VAERS [8]. The Australian Immunization Handbook states that vaccine providers should use their clinical judgment when deciding to report an event [9]. Consequently, the existing reporting system is potentially filtering out AEFI that individual practitioners decide not to report on. A study on Australian healthcare providers' knowledge and the challenges of AEFI reporting showed that reporting is infrequent and depends on their perception of what constitutes a reportable AEFI, with additional barriers of lack of time and knowledge about reporting processes [10].

Apart from potential underreporting by health professionals, Mesfin et al. point out that AEFI reporting that occurs *after* a patient has gone home following immunisation depends on patients or their caregivers returning to the clinic or visiting an emergency department or hospital [11]. Without such a visit, less severe AEFI are unlikely to be captured. Their study suggests that "AEFI-related calls" from telephone-based triage systems, offers opportunities for additional near real–time syndromic surveillance of AEFI, with the potential to identify severe AEFI signals earlier. In conclusion, alternate data source offer potential value to get a more timely and accurate picture of the quantity of possible adverse events.

Extensive use of social media has provided a platform for sharing and seeking health-related information. Social media data has consequently become a widely-used source of data for public health research [12]. In comparison with established traditional surveillance systems, social media monitoring is inexpensive, near to real time, and covers large populations [13]. Analysis of social media data can be used to supplement and corroborate established health reporting [14]. Examples of the use of social media for health reporting include communicable [15,16] and non-communicable disease [17,18] monitoring, assessing the impacts of health policies including vaccinations [19], drug use [20] and abuse studies [21], and pharmacovigilance [22] - which is the practice of the detection, assessment, understanding and prevention of prescription drugs adverse effects [23].

The primary aim of this study was to establish that social media monitoring for vaccine adverse event mentions was a viable enterprise, by applying NLP techniques to a relatively unfocused social media stream (tweets mentioning vaccines). The study also aimed to describe the NLP processes used, to assist other researchers to apply these to similar problems. These insights are formalized as the VAEM-Mine method, which encapsulates the workflow and techniques required to identify and combine the most effective of the processes for extracting vaccine adverse event mentions from general vaccine-related Twitter data. The proposed VAEM-Mine method is easily implementable and generally applicable to any similar problem of identifying personal health mentions based on the *type of language* used in them.

Background

Vaccine adverse event mentions

Vaccines belong to the broad category of medicines, in a subcategory known as "biologicals" [24].

Unlike medicines that are prescribed to limited populations as a course of *treatment* for a disease, vaccines are given to both healthy and vulnerable populations at large, sometimes over a short period, to enhance their immune systems' ability to combat a pathogen. In contrast to those who are taking a medicine to help to cure a disease or to treat unwanted symptoms, most people receiving a are not ill. Therefore, there is a deferred individual benefit to taking a vaccine, and consequently a very low acceptance of risk regarding vaccines [25]. Additionally, the pathophysiology of vaccine adverse events is not as well defined as those of adverse drug reactions, a reaction triggered by a vaccine could be caused by any of its multiple ingredients or even an error in administration [26]. Furthermore, a vaccine's "time to market" may be curtailed, such as has occurred in the COVID-19 pandemic and provide less opportunities for studying potential vaccine side effects over a large population for a long time.

Therefore, vaccines require a different emphasis in their safety surveillance. Monitoring for minor reactions is potentially just as important as surveillance for severe adverse events, as minor AEFI may act as a surrogate warning for more severe sequelae (such as increased rates of fever may be a marker for increased febrile seizures [11]), and also play a major role in affecting vaccine confidence [27]. Increased incidences of minor events could indicate larger problems and could ultimately affect public perception and acceptance of vaccines, and result in the failure of a vaccine program.

Vaccine safety surveillance systems' objectives are to monitor unexpected, rare and late-onset events *and* to observe changes in the rate of known and expected events. This research seeks to determine if social media monitoring can assist with the latter goal, because, as stated by Clothier et al. : "*While rare but particularly serious events can be detected through review of each individual report or active surveillance, an increased incidence in a more common AEFI is often more difficult to detect, and has been described as akin to 'finding a needle in the haystack'*"[28].

Monitoring of social media and user-generated data on the web enables timely and inexpensive gathering of much more information than can be accessed through traditional health reporting systems. The collective experiences and opinions shared by social media users, are an easily accessible wide-ranging data source for tracking emerging trends — which might be unavailable or less noticeable in data gathered by traditional reporting systems [28].

Social media monitoring for disease surveillance has been widely researched and proved useful in many areas including: tracking trends, early detection, forecasting, understanding transmission patterns, situational awareness, and discovering correlates of disease [13]. Vaccine-related social media monitoring offers the possibility of gaining early insights into vaccine safety issues through observing increased discussions by individuals experiencing vaccine reactions. The term "Vaccine Adverse Event Mention" (VAEM) is used in this research to refer to social media posts that *mention* vaccine adverse events, no matter their severity or specificity or proven association with a particular vaccine. This distinguishes VAEM from the signals used in previous research into the use of social media for vaccine and drug reaction surveillance, as these are looking for specific (and mostly severe) adverse vaccine events and drug reactions. This also distinguishes VAEM from formal AEFI reporting, as fewer barriers to reporting increase the chance that milder AEFI are described. VAEM are conversations, ideally gathered in volume, that contain information that might be those common AEFI that are so elusive to traditional reporting.

Pharmacovigilance using social media

Many researchers have successfully established the usefulness of social media as a pharmacovigilance source. Sarker et al. reviewed articles published from 2010 to 2014 on Adverse Drug Reaction (ADR) detection studies utilising social media [22]. They noted a shift of emphasis in this area of research, from exploratory studies to more structured approaches, which has resulted in an increased use of supervised machine learning techniques, and the need for annotated data. Their study highlighted the need for more annotated publicly available data for pharmacovigilance purposes. This led to their initiative of organizing shared tasks for Social Media Mining for Public Health Monitoring and Surveillance (SMM4H). The SMM4H shared task has been held annually

since 2016 and has always included tasks of binary classification of social media posts containing ADR, and of extraction and normalization of related terms [29–32]. Lardon et al. also conducted a scoping review to discover the extent of the use of social media for pharmacovigilance and concluded that more reliable pharmacovigilance data will be obtained as extraction systems mature, and that pharmacovigilance systems need to define the role that social media should play [33].

Vaccine monitoring using social media

There is a relative deficit in VAEM research, with investigations of vaccine and vaccination-related social media posts characterized as mostly concerned with sentiments, attitudes, and opinions. Salathé & Khandelwal analysed vaccine sentiments in Twitter posts about the influenza A (H1N1) vaccine [34]. Larson et al. (2013) found that vaccine-related subjects such as vaccine development and programmes were associated with neutral or positive sentiment, but that beliefs, perceptions, and issues of safety and vaccine impact were overwhelmingly associated with negative sentiment [35]. Another study assessed Twitter sentiments towards human papillomavirus (HPV) vaccines to understand public opinion and concerns [36].

Studies on using social media for ADR detection have included vaccine-related words in drug-related keyword searches used for collecting data from social media. An example is the work done by Sarker & Gonzalez [37], where 267,215 tweets containing 250 drug-related keywords, including “vaccine”, were downloaded over a period of four months. Smaller, cleaned, and labelled subsets of this corpus have also been published [38]. We downloaded and assessed these datasets; however, they did not contain any VAEM.

Wang et al. specifically addressed the challenge of influenza (flu) vaccine AEFI detection in posts by users that were known to have recently had a vaccination [39]. They emphasized that that the main problems for adverse events detection from social media were the cost of the annotation process and class imbalance. As a solution to the annotation problem, they based their work on first annotating users, then their tweets. To do this they needed to identify users who had vaccinations, and then collect all their subsequent tweets to look for VAEM. This approach still required tens of thousands of tweets to be annotated. For class imbalance they used a separate dataset of formal reports to add to the positive class. The emphasis of the study was to identify definite adverse events, and the language used in the formal reports was therefore able to contribute to the signal of what might be understood as an adverse event [40].

The VAEM-Mine method is dedicated to identifying vaccine adverse event mentions in a large corpus of Twitter data and is comprised of two main components. A topic modelling component to initially filter the data, as described in a prior publication [41], and a classification component to accurately identify the VAEM in the filtered data.

Materials and method

1. Data collection

The Twitter API was used with a search term of "vaccination, vaccinations, vaccine, vaccines, vax, vaxx, vaxine, vaccinated, vacinated, flushot, ‘flu shot’". No specific reaction mentions were used, instead the search concentrated on collecting vaccine-related posts likely to contain VAEM. Data was initially collected for five months, from 7th February to 7th June 2018, this was used for an initial training and evaluation of topic models and classifiers. Additional data was collected from 9th August 2018 through to 20th July 2019, which was used to verify the trained topic models and classifiers and to train more powerful classifiers. The resulting data consisted of a total of 811,010 tweets, and a daily average of 2,906 tweets. The data was split between the initial collection of around 400,000 tweets and the later collection of around 411,000 tweets. Preparing the data for processing reduced the data counts by removing duplicates and tweets that had no useful text. The final numbers were 328,822 from the initial collection, and 359,535 from the second collection – a total of 688,357.

Error: Reference source not found illustrates a sample of tweets that mention receiving vaccinations or vaccines. The first three examples contain genuine VAEM, but the others do not – even when the language is similar. Our goal was to first isolate the most likely records describing personal experiences of vaccination, then to refine that selection to those that are genuine adverse reaction mentions.

Table 1. Sample of vaccine-related tweets

aw wtf my poor arm is dead af from my flu shot
cannot lie on belly, baby gets squished; cannot lie on back, baby squishes; cannot lie on right side, i get heartburn; cannot lie on left side, vax arm is sore; let the third trimester moaning begin!
2 people recently, including my 88yo father, had flu shot and really bad reaction afterwards. both said it was probably as bad as getting the flu!!! flu2018 maybe undercooked the vaccine
I got vaccinated as a kid. As a result, I'm now starting to gray and bald. My balding got so bad I had to shave my head. I've also gained weight. Because of vaccines I've started aging instead of dying as a baby.
Urgent vaccination plea after measles outbreak in West Yorkshire
Researchers are developing a personalized vaccine which they hope could tackle ovarian cancer

The topic modelling showed that VAEM and similar personal health mentions were a distinct topic, and therefore that topic models could be utilised to filter for the tweets that were most like VAEM, so that more homogenous and concise datasets could be created, for labelling and subsequent training of classifiers. Filtering data for classification via topic modelling was adopted as a core component of our approach, which we call the VAEM-Mine method. A previous publication [41] described the process of choosing the best performing topic models for the method, including a detailed description of the scoring method used. A later section describes the VAEM-Mine method.

2. Classification

This paper focuses on analysing classifier performance in relation to the available data. As described in the previous section, data was collected in two phases. Topic models were trained on the first phase data and used to filter that data and the subsequent second phase data into likely VAEM-containing datasets, which were then used for classification. Classifiers were trained and assessed with the filtered first phase dataset, and the combined (filtered) first and second phase datasets. The following section describes the creation of these datasets; the subsequent section describes the classifiers.

Classification Datasets

After topic model-based filtering was used, the original prepared data collections of 328,822 and 359,535 tweets were reduced to more VAEM-like datasets of 18,801 and 80,372 tweets. Therefore, filtering eliminated around 85% of the data, which did not contain any significant numbers of VAEM. These more VAEM-focused datasets were binary labelled to train classifiers, as either VAEM and non-VAEM. Although only 10.2% of the tweets were identified as VAEM, this was a considerably better proportion of VAEM compared to the original data, which we estimated by sampling to contain VAEM in only 1.5% of the tweets.

Balanced datasets of 3,519 and 15,730 tweets were created from these imbalanced datasets, together with hold-out test datasets – these were an imbalanced test set of 614 tweets, and a balanced test set of 828. The main datasets were named the “Phase One” and “Phase Two” datasets, and the test

datasets were referred to as the “Phase-One Test” and the “Phase-Two Test” datasets.
The imbalanced Phase-One Test dataset of 614 tweets came from Victoria, Australia in the lead up to and during the 2018 flu immunization period. These tweets were assembled to enable comparison of tweet trends in our local situation with statistics from the Victorian vaccine authority SAEFVIC. With 90 VAEM and 524 non-VAEM, the test set is quite imbalanced, but reflects how the data came through the topic model-filtering process, without any subsequent balancing. The Phase-One Test dataset was used as a benchmark throughout the classification testing, though we noted that classifiers identified as scoring best with this data did not necessarily also perform as well with the larger Phase-Two Test dataset.
The datasets were combined to re-train classifiers, and to train larger Transformer-based classifiers – becoming a “Combined” dataset of 19,249 tweets, and a “Combined Test” dataset of 1,442 tweets. The training data was split into training and validation data with a 75/25 ratio.
See Appendix A for the description of how the datasets were constructed.

Classifiers

We identify the classifiers used as either traditional models or neural networks. These categories refer to the data preparation used as well as the underlying architecture. Our default data approach with standard classifiers was “bag-of-words” [42] represented via compressed sparse matrices. We used SKLearn [43] vectorizing libraries such as TfidfTransformer [44] libraries for tokenizing lower-case text for the standard classifiers. A grid or random search was used to ascertain the best combinations of vectorizer, stop words and numbers removal, and n-grams. The neural networks used dense word embedding vectors via a Word2Vec Skip-gram corpus [45] for CNNs and LSTMs, and we built the Word2Vec corpus using Gensim library functions [46] utilizing all of the Twitter data we had collected. The Transformer models used byte-pair-encoding (BPE) [47]; the BPE tokens were derived just from the filtered texts we had retained from topic modelling. The classifiers are listed in Error: Reference source not found.
Table 2. List of classifiers

Models	**Library / Github source**
Logistic Regression CV	sklearn.linear_model
Stochastic Gradient Descent Classifier	sklearn.linear_model
Linear Support Vector Classification (SVC)	sklearn.svm.SVC
Random Forest Classifier	sklearn.ensemble
Extra Trees Classifier	sklearn.ensemble
Multinomial Naïve Bayes	sklearn.naive_bayes
Naïve Bayes SVM (combined NB and Linear SVM)	GitHub Joshua-Chin/nbsvm
XGBoost	GitHub dmlc/xgboost
Ensemble (Naive Bayes SVM, Logistic Regression CV, SGD Classifier, Linear SVC, Random Forest)	Using majority voting on predictions
CNN, LSTM, BiLSTM,GRU, BiGRU, CNN-BiLSTM, CNN-BiGRU	Pytorch; RaRe-Technologies/genism; Shawn1993/cnn-text-classification-pytorch; bamtercelboo/cnn-lstm-bilstm-deepcnn-clstm-in-pytorch
RoBERTa, RoBERTa Large, BERT,XLNet, XLNet Large, XLM	Pytorch; huggingface/transformers

3. VAEM-Mine method

The classification models we analyse in this paper were the final component of a pipeline of processes, starting with data collection, then by cleaning and processing through topic models to filter for data that was as close as possible to the VAEM we were trying to identify with the classifiers. This allowed for a focussed binary classification approach for isolating VAEM. We describe this pipeline as the VAEM-Mine method (Figure 1). The use of “Mine” in the method name reflects the process involved in detecting VAEM in Twitter conversations – the raw material in the form of Twitter texts must first be collected and prepared, then filtered to *extract* the valued VAEM-containing data, like the processes involved when mining for ores. We summarize the pipeline here to give context to the classification component, and to allow us to describe the overall effectiveness of the method.

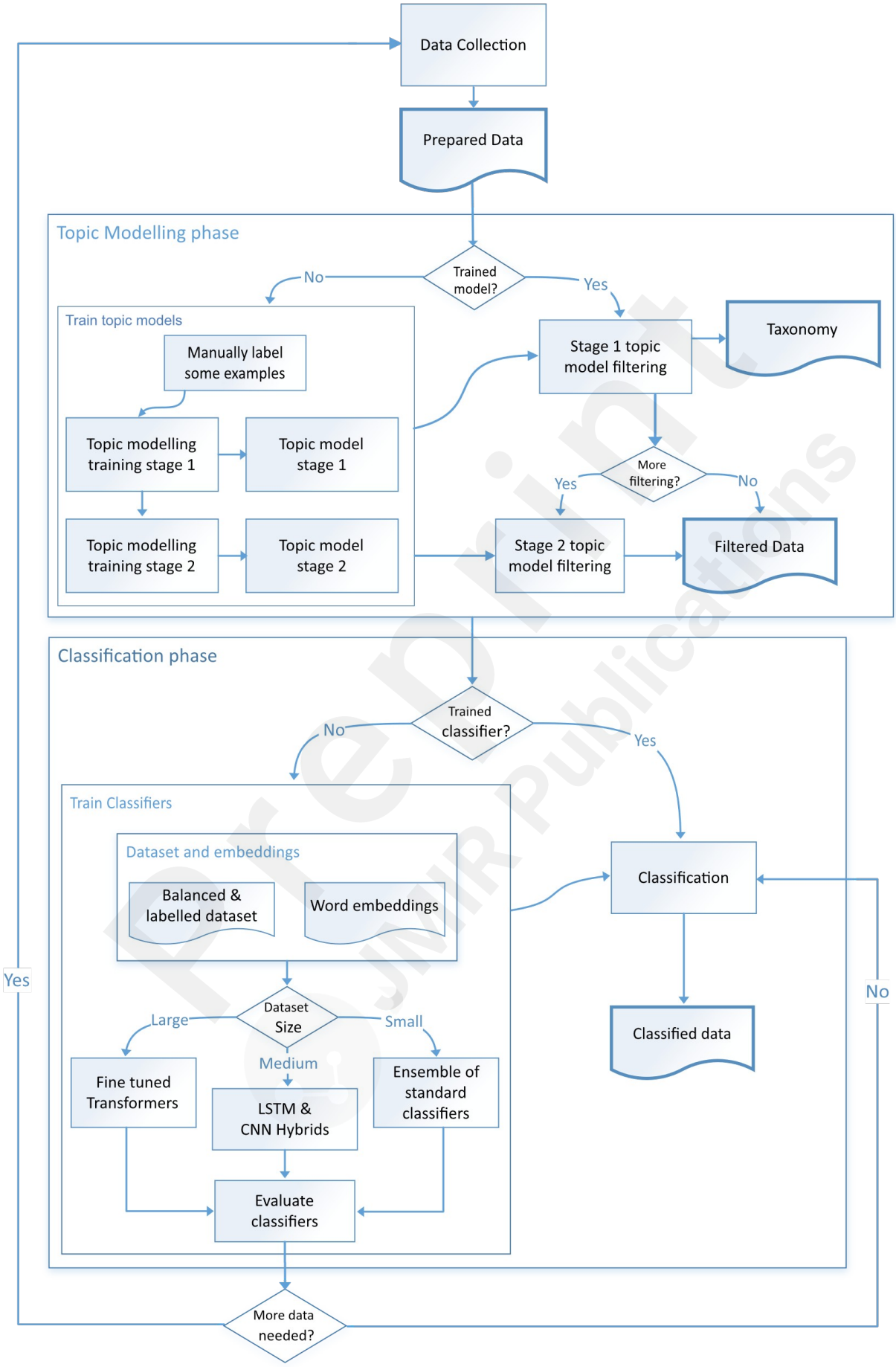


Figure 1: The VAEM-Mine method

The method includes decision points to determine the appropriate direction, either the training process, or the application of the trained models to incoming data. When the topic modelling phase is entered for the first time, a trained model does not exist, so the work of training the topic models begins. The first step is to label some examples of the subject of interest (in this case VAEM), and additional examples of other subjects. This enables the application of a topic modelling scoring approach that measures how the VAEM-label of interest is distributed in the topics, compared with other labelled topics. A topic model is considered to score well if the VAEM are concentrated in just a few topics, and ideally in just one topic, without also including too many of the other labels. Further refinement of the data is possible by a second stage of topic modelling on the data obtained from the top model of the first stage. The second stage identifies topics that have a higher ratio of the subject of interest to other subjects in the texts, but at the expense of losing some texts containing the subject of interest. Having trained the models, they can be applied to filter the incoming data, and it is up to the user whether they take just the output of the best topic(s) of the first-stage topic model, or to further refine the data by taking it from selected topics of the second-stage topic model. The topics of the first stage of topic modelling are also potentially useful to obtain a domain taxonomy.
The filtered data is handled by the classification phase, which also has the decision point for either training classifiers or using trained classifiers. When training, the choice of classifiers should relate to the quantity of available data, and if results are not as expected then a decision may be made to obtain more data. The method requires the incoming filtered data to be labelled for the creation of datasets suitable to train the classifiers. It additionally requires the creation of domain-specific embeddings.

Results

1. Classification Analysis

Classification training and evaluation was conducted twice, firstly with the filtered data that was obtained from applying topic modelling to the initial phase of data collection, then with data obtained through topic model filtering over all the collected data. The following sections describe these as Phase-One and Phase-Two classification.

Phase-One classification

The first phase of classification experiments used a training set of 2,639 records, a validation set of 880 records, and the imbalanced holdout Phase-One Test dataset of 614 tweets. The F1 Scores for the models evaluated in this phase are listed in Error: Reference source not found.

Table 3. Phase-One F1 Scores

Model	Validation	Imbalanced Test	Balanced Test	Combined Test
CNN - BiGRU	**0.842**	0.762	**0.846**	**0.825**
BERT	N/A	0.767	0.841	0.824
BiGRU	0.807	0.793	0.828	0.822
CNN - LSTM	0.805	0.777	0.815	0.808
BiLSTM	*0.815*	**0.807**	0.807	0.807
GRU	0.820	0.730	0.822	0.804
CNN - BiLSTM	0.816	0.766	0.810	0.802
CNN	0.816	0.787	0.800	0.798
LSTM	0.796	0.767	0.803	0.796
Ensemble	0.815	0.726	**0.829**	**0.810**
Logistic Regression CV	0.812	0.730	0.820	0.803
Linear SVC	0.814	0.693	0.824	0.797
Stochastic GD	0.805	0.636	0.825	0.785
Naïve Bayes SVM	0.792	**0.767**	0.789	0.785
Random Forest	0.814	0.694	0.801	0.779
Extra Trees	**0.833**	0.688	0.801	0.777
XGBoost	0.811	0.704	0.791	0.774
Naïve Bayes	0.798	0.605	0.799	0.756

Table 3 includes subsequent tests of the models against the later Phase-Two “Balanced test” dataset and a “Combined test” dataset that uses all the test data. F1 Scores are measured for the positive, VAEM class, rather than over both classes. The models are arranged in order of the best F1 Score over the test datasets; validation scores are also included, where available. There are no validation F1-Scores available for models using transfer learning — they used a cross-validation approach and so were given combined training and validation data and were evaluated only against test datasets.

The Ensemble model is shown in the middle of the table, which was scored based on a maximum voting of the predictions of 5 traditional models on the test dataset, it had the overall best score on the larger test data when using standard classifiers, which are all arranged below it. Phase-One classification was completed with the assessment of the BERT Transformer model, which did not perform as expected.

All the deep learning models outperformed the best traditional classifier on the Imbalanced test dataset, by at least 6% and almost as much as 10% - the improvement was mostly due to a greater capacity to correctly distinguish non-VAEM-related tweets, and so obtain a greater precision. However, when evaluated against the Balanced and Combined test sets the results differed — here the traditional classifiers outperformed many of the deep learning models, especially the Ensemble, which was only surpassed by the top 3 deep learning models.

The best of all other experiments with CNNs placed them below two CNN combined models — one combined with a bi-directional GRU (CNN-BiGRU), the other combined with a bi-directional LSTM (CNN-BiLSTM). It is notable that the bi-directional versions of these models generally outperformed their standard counterparts.

Phase-Two classification

The second phase of classification used five times as many records to train the models, by combining

the 3,519 training records from the first phase with another 15,730 records, resulting in a total of 19,249, and by introducing the Phase-Two Test dataset of 828 records. The greater amount of data allowed a proper assessment of neural networks, but it also improved model performance across the board – see Error: Reference source not found. The “Imbalanced Change” and “Combined Change” columns shows the percentage increase of the models’ F1-Score over the Imbalanced Test and Combined Test datasets, compared to their Phase One equivalents.

There was a much greater consistency of scoring over all the test datasets, and the top models scored best over all the test datasets. The highest score was from the RoBERTa Large Transformer model, with an F1 of 0.919 on the Imbalanced data, the standard RoBERTa model was placed second.

Table 4. Phase-Two F1 Scores

Model	**Validation**	**Imbalanced Test**	**Balanced Test**	**Combined Test**	**Imbalanced Change**	**Combined Change**
RoBERTa Large	N/A	**0.919**	**0.908**	**0.910**	-	-
RoBERTa	-	0.901	0.905	0.904	-	-
XLNet Large	-	0.884	0.906	0.902	-	-
XLNet	-	0.870	0.903	0.897	-	-
XLM	-	0.910	0.894	0.897	-	-
BERT	-	0.863	0.892	0.887	12.6%	7.7%
BiGRU	0.877	0.855	0.896	0.890	7.9%	8.2%
CNN-BiGRU	0.874	0.849	0.890	0.884	11.4%	7.1%
LSTM	0.866	0.875	0.879	0.878	14.1%	10.3%
CNN-LSTM	0.866	0.862	0.876	0.873	10.9%	8.1%
BiLSTM	0.872	0.847	0.884	0.878	5.0%	8.8%
GRU	0.869	0.825	0.876	0.868	13.1%	7.9%
CNN-BiLSTM	0.872	0.824	0.879	0.871	7.6%	8.6%
CNN	0.864	0.805	0.866	0.856	2.4%	7.2%
Ensemble	**0.870**	**0.818**	**0.874**	**0.865**	12.6%	6.8%
Logistic RCV	0.866	0.807	0.873	0.861	10.5%	7.3%
Stochastic GD	0.865	0.806	0.873	0.861	26.7%	9.7%
Linear SVC	0.864	0.802	0.869	0.857	15.7%	7.5%
Random Forest	0.857	0.796	0.864	0.853	14.7%	9.5%
Extra Trees	0.857	0.789	0.862	0.849	14.7%	9.2%
NB SVM	0.838	0.798	0.838	0.832	3.9%	5.9%
XGBoost	0.845	0.714	0.854	0.831	1.3%	7.4%
Naïve Bayes	0.835	0.735	0.841	0.822	21.5%	8.7%

One of the most noteworthy effects of having more data is that the previously strong combinations of CNN with BiGRU and BiLSTM models were surpassed by the LSTM on the Imbalanced test data, both when combined with a CNN but most significantly as a stand-alone model. The LSTM in fifth position on the Imbalanced Test scoring is only 2.5% behind the score of the RoBERTa Large model. One can fairly conclude that a CNN or hybrid CNN approach performs well when limited data is available but will likely be surpassed by architectures designed for sequential language processing as more data becomes available.

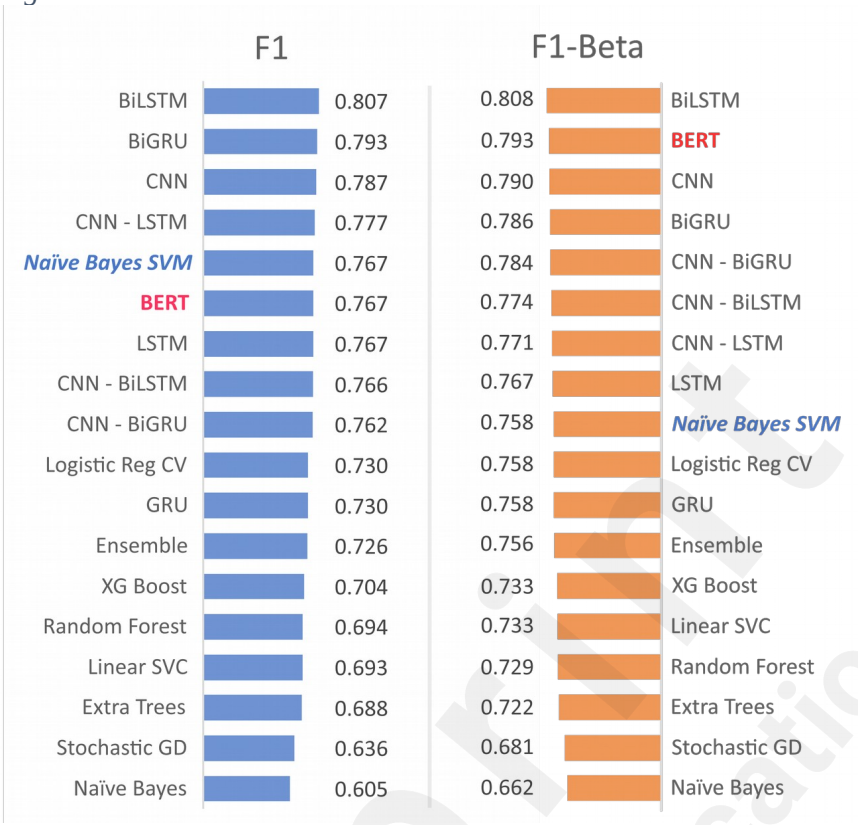
Classification performance analysis

Our initial classification training and evaluation was conducted after we had collected enough data to perform the topic model training process, but while we were processing this first phase, we continued to collect data. Our assessment of the models trained on the initially collected data indicated that they should improve with more data – we have included a decision point for this model assessment and possible re-training into the VAEM-Mine method. The results depicted in the F1 Scores tables show the expected improvements after training with more data. The following analysis examines how the models performed with these two training phases – we believe this can be useful for anyone else who is dealing with similar types and volumes of data, when trying to decide on the most appropriate model, or whether they should continue to collect data.

Imbalanced Test data with Phase-One models

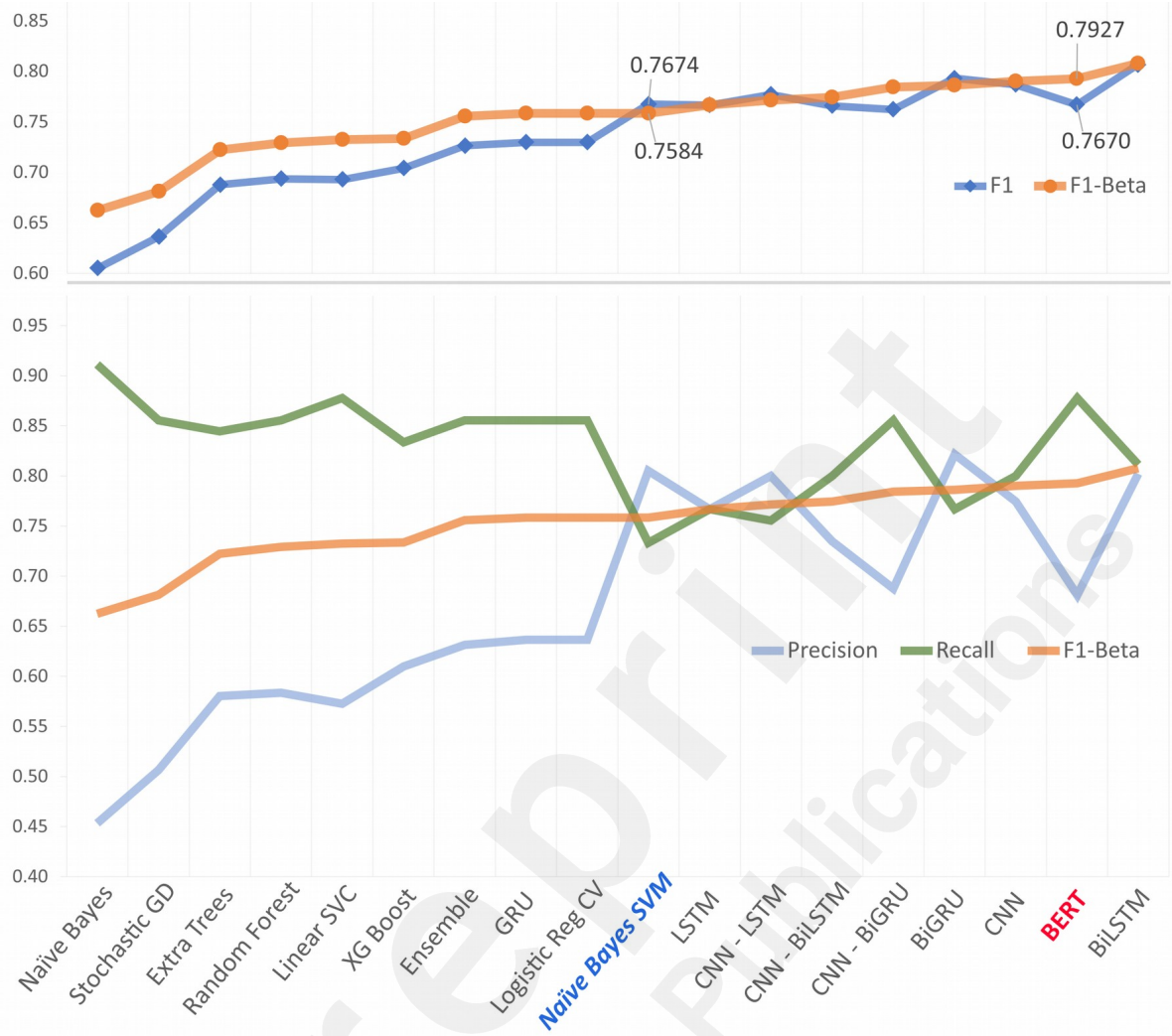
The Phase-One Test dataset was highly imbalanced, having only 90 VAEM vs. 524 non-VAEM – but it was used throughout as a standard because it represented a real scenario of tweets that had been identified as belonging to our geographical region, and that had been processed by topic modelling but had not been subsequently balanced. During our initial, Phase-One testing, we noted that the Imbalanced Test dataset suited models that favoured the non-VAEM (negative) class. That is, models that tended to shift both false and true positives into the negative class did well with this test set. For instance, the Naïve Bayes SVM model eliminated many false positives and thereby achieved the highest precision among the traditional models, but it also eliminated true positives and so had the poorest recall. However, it was awarded the highest F1 Score among the traditional models just because there was a much high number of the non-VAEM class in the test data. That is, the model's precision benefitted by removing many false positives, which offset the penalty due to its removing (somewhat fewer) true positives. Because our requirement was to identify as much VAEM as possible with relative precision, we needed to favour VAEM recall over precision; therefore, we also tested with an F1-Beta score, using a beta of 1.3. We observed that the Naïve Bayes SVM F1-Beta score was then closer to the middle of the scoring range, and that BERT, with its relatively higher recall, was promoted to second place – see Error: Reference source not found.

Figure 2. Phase-One models and Imbalanced Test data - F1 vs F1 Beta



The relationships between F1 and F1-Beta, together with the variations of precision and recall that help to explain the model's performance, are depicted in Error: Reference source not found. The models are displayed in increasing order of their F1-Beta scores. Note that Naïve Bayes SVM has one of the highest precision values, but also the lowest recall, and so is penalised by the F1-Beta score. Conversely, the BERT model has a high recall but a relatively lower precision, resulting in the same F1 Score (0.767) as the Naïve Bayes SVM model – but that because of its recall is favoured by the F1-Beta score. The chart shows that the traditional classifiers tend to have a higher recall but with poor precision, but that after the point where the Naïve Bayes SVM enters the chart, the remainder of the models (which are based on neural networks) have precision and recall values that are somewhat closer to each other.

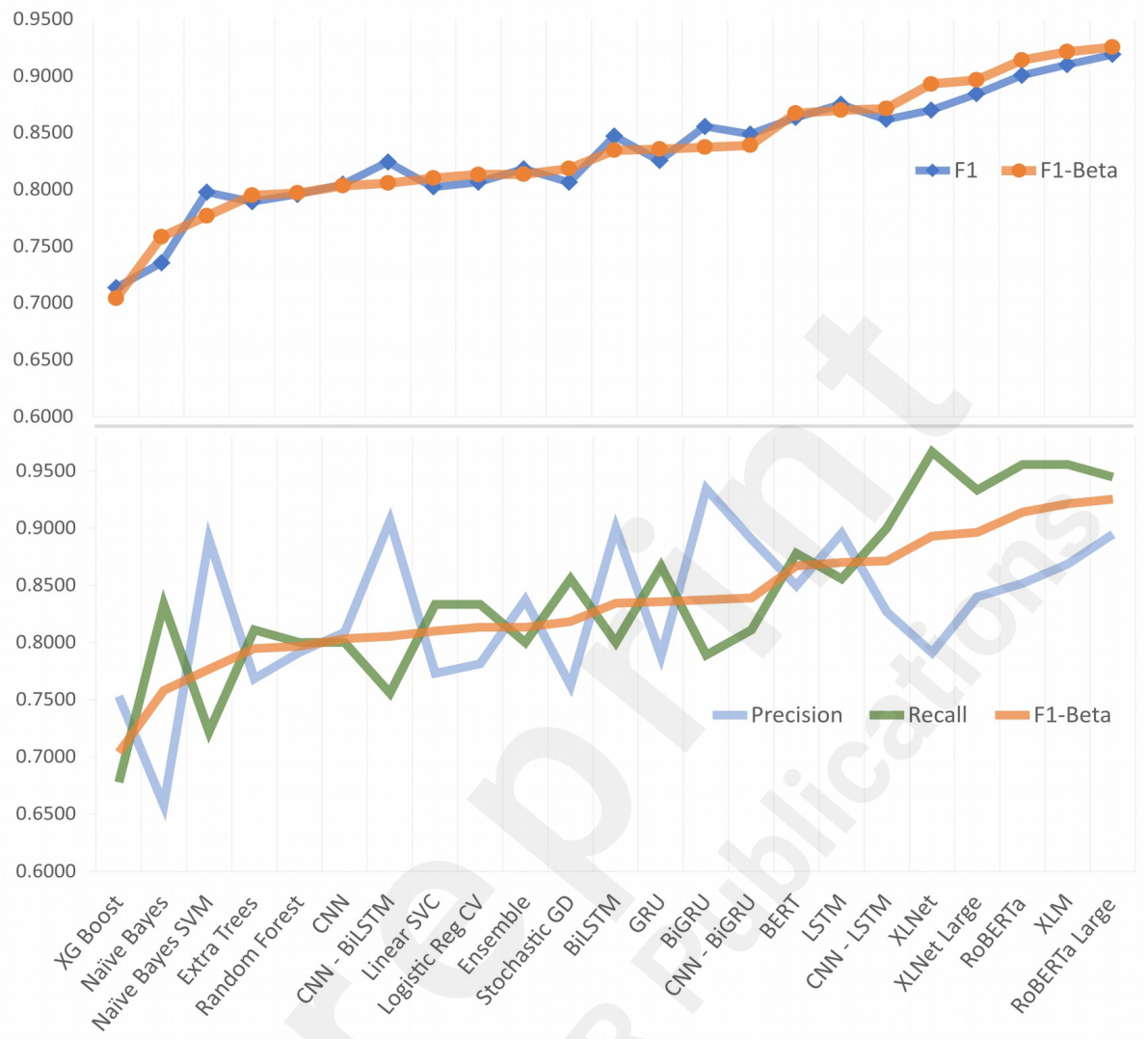
Figure 3. Phase-One models on Imbalanced Test data - F1 Scores and their measures



Imbalanced Test data with Phase-Two models

Almost all the models' F1 scores increased by at least 5%, after the models were re-trained with the additional data available in the second phase of training - where the training data increased by over 5 times from 3,519 records to 19,249. The Phase-Two classifiers included five new Transformer models. See Error: Reference source not found for the performance measures against the Imbalanced Test data, which again show some extreme differences in precision and recall, but significantly, fewer examples of a great divergence between precision and recall among the lower-order models, with F1 and F1-Beta being more aligned. The upwards trajectory of the resulting F1 scores is rather steeper than it was with the Phase-One models – there is a 20% difference between the worst and best performing models. This is due to the top performing Transformer models having a 10% better performance than the top performing traditional models, and effectively coping with the imbalanced data – the "Balanced Test data with Phase-Two models" section shows that they were able to obtain very similar F1 scores on this data as they did on the balanced Test data. The RoBERTa Large model achieved an F1 score of 0.919, and an F1-Beta of 0.923.

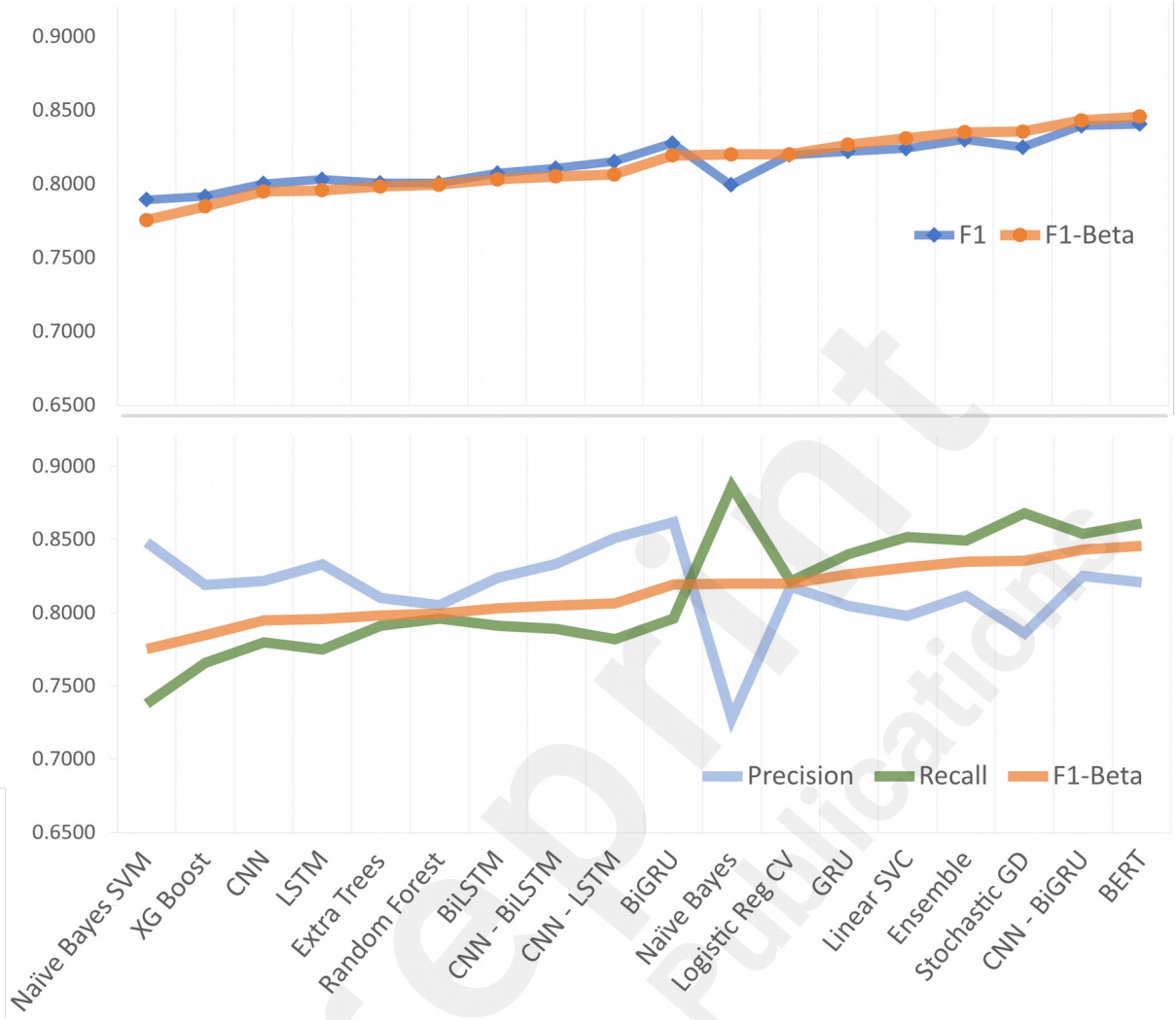
Figure 4. Phase-Two models on Imbalanced Test data - F1 Scores and their measures



Balanced Test data with Phase-One models

The Balanced Test dataset consisted of 828 records with 431 VAEM and 397 non-VAEM. The behaviour of the models when tested with this data was a lot more regular, even with the Phase-One models – see Error: Reference source not found.

Figure 5. Phase-One models on Balanced Test data - F1 Scores and their measures

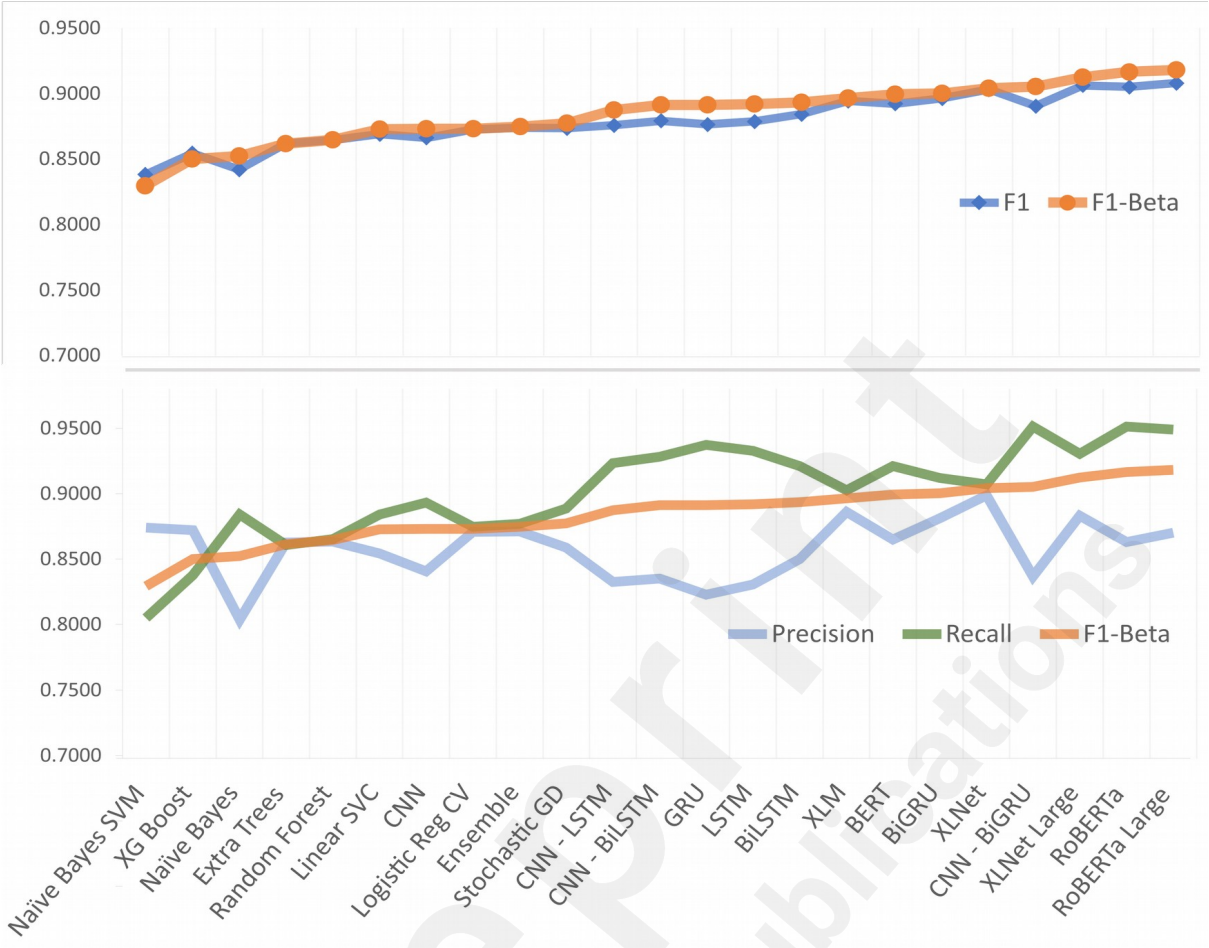


Precision initially exceeded recall, then switched to recall exceeding precision, but with a more consistent relationship when compared with the evaluations on the Imbalanced Test data. The Naïve Bayes-based models are noteworthy: In this scenario Naïve Bayes SVM was the poorest performer, and standard Naïve Bayes scored much better – its very high recall was weighted by the F1-Beta calculation to offset its poor precision and give it a score in the middle of the range. BERT and CNN-BiGRU were the best models – they both had combinations of high recall and precision, but the Ensemble of traditional models had a similar balance of precision and recall and was the fourth best model. Testing on the Balanced Test data did not show such a clear distinction between the performance of the traditional models and neural networks, several traditional classifiers performed better than some of the neural networks.

Balanced Test data with Phase-Two models

As previously noted, the models' F1 scores increased by at least 5% after the models were re-trained with the larger data, and when tested with the balanced Test data their performance improved compared to that when tested with the Imbalanced Test dataset, all the F1 scores were above 0.8 – see Error: Reference source not found.

Figure 6. Phase-Two models on Balanced Test data - F1 Scores and their measures

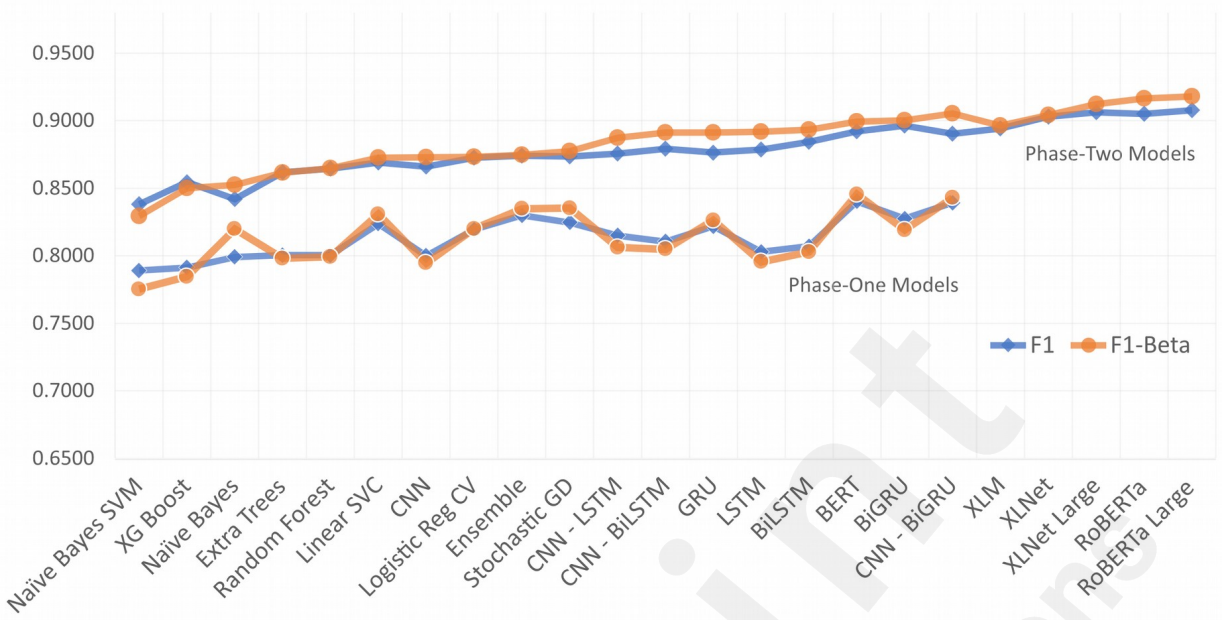


The Phase-Two results depicted in Error: Reference source not found show an even closer relationship between the F1 and F1-Beta scores, with a clear progression in scores from the traditional models to the neural network-based models – there are no longer traditional models' results interspersed within those from the neural networks models. The RoBERTa Large model achieved an F1 score of 0.908 and an F1-Beta of 0.918 – as indicated earlier these align with the scores achieved on the Imbalanced Test data, so we conclude that training on the larger dataset has achieved an optimal result from the model.

Phase-Two models vs Phase-One models

Error: Reference source not found shows the relative performance of the Phase-Two models vs the Phase-One models when evaluated with the Balanced Test data – as such they represent the best performances of both training phases. There are five extra entries for the Transformer models added to the Phase-Two models. Note that the highest scores from the Phase-One trained models shown in the bottom part of the chart are around 0.85, whereas only the bottom three Phase-Two trained models are on or below 0.85; that almost all scores have increased by at least 5%; and that there is a greater overall rate of improvement noticeable in the slope of the Phase-Two trained models.

Figure 7. Phase-Two vs Phase-One models - F1 scores on Balanced Test data



To conclude this section, the descriptions of classification results and the analysis presented above should assist other researchers to make decisions about classifiers and the amount of data required, should they be encountering similar texts and volumes of data. Our method's design incorporates an evaluation of the classification process with the decision point that allows for collection of more data. Our evaluation of the models on the 3.5 thousand record dataset indicated that more data could improve the models. After retraining the models on 19 thousand records and evaluating with a larger balanced test dataset we have seen significant performance gains and the ability to use more powerful classifiers, with the Transformer models proving to be most capable. The F1 scores of the Roberta Large model are exemplary and have established that classification very effectively isolates the VAEM from the incoming data, which is the output of the topic modelling component of the VAEM-Mine method. The next section assesses the end-to-end performance of the VAEM-Mine method for the task of identifying and isolating vaccine adverse event mentions from the vast amount of other vaccine-related Twitter conversations.

2. VAEM-Mine Method Performance

The VAEM-Mine method is a combination of data preparation, then a two-stage topic modelling phase, and finally a binary classification phase. The previous section has analysed the classification phase of the method, and a prior paper has analysed the topic modelling phase [41]. Here we assess the overall effectiveness of the method, in terms of the quantities of tweets having vaccine adverse event mentions that were progressively filtered out by the method. The values presented are the total numbers of data collected and processed via the method, with estimates where appropriate.

Error: Reference source not found depicts the numbers obtained after data collection through to the completion of the topic modelling. After cleaning the data, topic modelling was used to process 688,357 records. Stage One of topic modelling filtered out 570,383 records to retain 117,974 records likely to contain VAEM. The data was around 14.5% of the original total, and contained over 99.5% of all available VAEM.

Table 5. Summary Topic Modelling counts

Steps	Counts	Percentages
Tweets Collected	811,010	
- Cleaned	-122,653	
- Discarded (Stage One)	-570,383	
Tweets after Stage One	117,974	14.5% of initial data
- Discarded (Stage Two)	-19,083	
Tweets after Stage Two	98,891	12.2% of initial data
Stage Two proportions		
Non-VAEM	88,900	
VAEM	9,991	10.1% of Stage Two data 1.2% of initial data
VAEM proportions		
In other Stage Two topics	2,367	
In best Stage Two topic	7,624	76.3% of VAEM

To prepare for the first round of classification, 19,083 records were discarded – that is all records that were not in the top three topics of the Stage Two topic model. Subsequent labelling showed that the discarded data also removed 94 VAEM from the data, which was approximately 5% of the VAEM in the first round. For the second round of classification all the records identified as likely VAEM by the first stage topic model were retained. The resulting 98,891 records over both rounds of classification were labelled, and VAEM were found to be 10.1% of the retained data, constituting 9,991 posts. The Stage Two topic models' topic numbers were assessed, and it was found that the best Stage Two topic contained 7,624 VAEM, which was 76.3% of the retained VAEM, and there were around 10% more VAEM than non-VAEM in the topic.

We conclude from these figures that topic modelling is an effective filtering mechanism, as it identified virtually all the VAEM while removing a lot of unwanted data. The filtered data was more manageable for labelling for classification than it would have otherwise been, and if needed the filtered output of the Stage Two topic model could be used as it is, with the understanding it discards some VAEM, and still contains a smaller but similar number of non-VAEM. But, as discussed above, classification is a more precise final step to obtain VAEM from the filtered records.

Classification phase

To assess classifier effectiveness in relation to the total data, the recall and precision of the best classifier, the RoBERTa Large model, were applied to the total VAEM to obtain an *estimate* of its performance on the total VAEM. These were a precision of 0.874 and a recall of 0.948 on the combined test data.

- Applying the recall to the total 10,085 VAEM-containing tweets, we estimate that 9,562 (94.5%) of the VAEM tweets would be correctly classified, missing 523 (5.5%) of the VAEM.
- By applying precision, we estimate that 1,374 (1.4%) of the non-VAEM would be included in the data.
- These results of 94.5% of VAEM together with 1.4% of the non-VAEM in the predicted positive class are clearly superior to those obtained with the best topic of Stage Two topic modelling, where we saw the proportion of VAEM in the best topic was 76.3% and where the almost equal number of non-VAEM in the topic was around 7.2% of the non-VAEM.

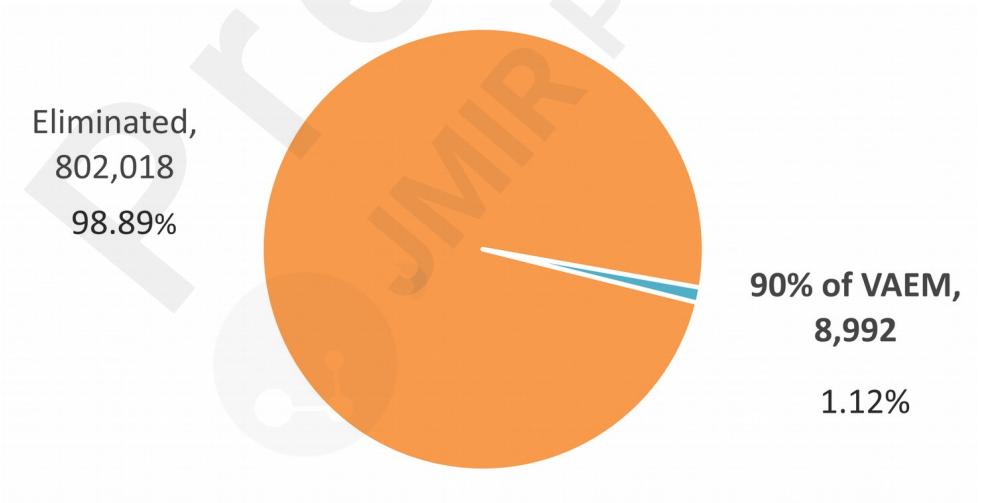
Combined topic modelling and classification effectiveness

Measuring the combined effectiveness of topic modelling and classification:

- 8,992 VAEM are identified from the original 811,010 records, conservatively estimated as being at least 90% of all likely VAEM.
- 802,018 non- VAEM are eliminated through cleaning, topic modelling, and classification.
- 10% of the VAEM are also eliminated in this grouping, the attrition is a consequence of the filtering and classification required to capture the 90%.
- In overall percentage terms, 98.89% of data is eliminated as *not* containing VAEM, with a very small amount misidentified, to identify 1.12% of the data as having VAEM, with a 90% success.

Error: Reference source not found shows the estimated combined effect of the topic modelling and classification processes of the VAEM-Mine method.

Figure 8. VAEM-Mine method - Capturing of 90% of VAEM



The results indicate that the combined approach of topic modelling followed by classification effectively identifies and isolates vaccine adverse event mentions from almost all other vaccine-related Twitter posts. The VAEM-Mine method enables us to identify the most effective topic models and classifiers for the core task of isolating VAEM. In particular, the key to the method's success is the topic modelling phase which drastically reduces the amount of irrelevant data and so delivers manageable data to the classification phase. As NLP technologies improve and new topic models and classifiers can be introduced, then we assume that even these results will improve.

Discussions

1. Principal Findings

The key objective of this study was to contribute to research on vaccine safety surveillance, by illustrating that social media monitoring has potential to augment existing surveillance systems. We have demonstrated a near-real time capable automated topic modelling and classification method for identifying VAEM with a high degree of sensitivity and specificity following vaccination. The method approached the problem of finding sparse vaccine adverse event mentions by first using topic modelling to focus on the semantic nature of such posts, which were predominantly those describing personal health issues in relation to vaccines. The subsequent data, filtered by using topic models, simplified the tasks of labelling and of training classifiers to identify vaccine adverse event mentions with a high degree of accuracy. We have demonstrated that standard NLP topic models and classifiers can be combined to isolate VAEM, based largely on the type of *language* used when describing adverse events.

The approach previously described by Wang et al.(2019) to detect influenza vaccine AEFI from Twitter, required purpose-built classifiers that were looking for specific adverse event reactions from tweets belonging to an already identified subset of Twitter users, who were known to have recently received a flu vaccine. Their classifiers were trained to identify known reaction keywords derived from a medical database, our approach relies on semantic features of the tweets to elicit the likely cohort, and the power of modern Transformer classifiers to determine the true signals. In short, by tackling the problem of finding adverse events through the lens of the language used in personal health mentions, we have been able to prove that social media can provide of wealth of useful data – without requiring any intensive labelling and specific training of NLP models to recognize adverse event words.

This study described the method used to identify vaccine adverse event mentions, which has two phases, a topic modelling process followed by classification. The most effective topic models were determined using F1-scoring over a small number of labelled posts. The scoring approach was a key part of the success of the topic modelling and worked by ascertaining when topic models were most effective at including VAEM into one topic. Identifying the effectiveness of the topic modelling scoring approach, and the techniques employed to use it, are important contributions of the research.

The VAEM-Mine method has a significant capability by to successively isolate vaccine adverse event mentions from the massive amount of other vaccine-related Twitter posts. The topic modelling phase was able to isolate up to 99% of the Twitter posts which contained VAEM. This was just 1.1% of the original data, thereby eliminating 98.9% of irrelevant posts. A second stage of topic modelling proved to be effective at further isolating VAEM from this dataset, but ultimately the classification phase could identify VAEM with an F1-Score of 0.91.

This research also presented a detailed reporting and comparisons on a range of classification models, including traditional machine learning models and deep neural (deep learning) networks. Their effectiveness was measured against different sized datasets, emulating data sizes that are likely to be available to other researchers. Therefore, insights into relative model effectiveness vs data size will be useful to other researchers wanting to use commonly available techniques. The research observes that the most powerful deep learning models only excel when given more data, which is well known, but the research quantifies and compares results to give a concrete understanding about the data needs of the range of models.

There are unavoidable issues and potential biases that result from using any social media data. One limitation of this study is the use of only Twitter as a data source, with validation using other social media data sources needed to maximise sampling within given populations and across different nations. While the data collection for this spanned a year, and included some potential trend patterns during influenza seasons, a longer-term data collection would be better for this kind of analysis. The

full year's data was required to properly train and evaluate the classifiers.
The proposed VAEM-Mine method is comprehensive, efficient, easily implementable, and potentially applicable to any similar problem of identifying personal health mentions based on the types of language used in them. The specific technique identified in this research of F1-scoring based on a small number of labelled posts is a practical and easily implementable solution.
The techniques used are applicable to any social media platform. The research confirms that social media can become a valuable complementary source for vaccine safety signal monitoring.

Conclusion and future research

We have determined that the VAEM-Mine method is an effective approach for both identifying and applying the topic models and classifiers that, when combined, can filter out the vast amount of irrelevant vaccine-related conversations and isolate vaccine adverse event mentions. In our previous papers we have described and assessed the topic modelling component of the VAEM-Mine method that is used for the initial filtering of data, here we have examined the classification component and the overall effectiveness of the method. We are confident that the VAEM-Mine method is suitable for the task of targeted filtering of social media messages for similar distinctive discussions.
The nature of the language in VAEM social media posts is reasonably consistent despite the variety of terms used. The use of topic modelling to encapsulate these similar posts into one topic, is adaptable to any similar problem. For instance, social media posts concerning the current COVID-19 pandemic are immense, but those that are concerned with personally experiencing the virus or vaccines are miniscule in comparison, yet they contain similar language. These techniques could be applied to help to isolate those posts now, as well as being applicable later to help identify trends in any developing health crises in relation to the new vaccines. A key finding of the research is that appropriately scored topic modelling is highly effective for identifying social posts that might contain VAEM. The specific technique identified in this research of F1-scoring based on a small number of labelled posts is a practical and easily implementable solution.
Acknowledgment: The authors would like to thank Mr. Christopher Palmer for providing technical advice in the project.
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Conflict of Interest: The authors declare that they have no conflict of interest.
Ethical approval: Ethics approval for this study was granted by Monash University Human Research Ethics Committee (Project ID: 11767).

References

[1] Chen RT. Vaccine risks: Real, perceived and unknown. Vaccine 1999;17:41–6. https://doi.org/10.1016/S0264-410X(99)00292-3.
[2] Jacobson RM, Adegbenro A, Pankratz VS, Poland GA. Adverse events and vaccination-the lack of power and predictability of infrequent events in pre-licensure study. Vaccine 2001;19:2428–33. https://doi.org/10.1016/S0264-410X(00)00467-9.
[3] Griffin MR, Braun MM, Bart KJ. What should an ideal vaccine postlicensure safety system be? Am J Public Health 2009;99:345–50. https://doi.org/10.2105/AJPH.2008.143081.
[4] Chen RT, Shimabukuro TT, Martin DB, Zuber PLF, Weibel DM, Sturkenboom M. Enhancing vaccine safety capacity globally: A lifecycle perspective. Vaccine 2015;33:D46–54. https://doi.org/10.1016/j.vaccine.2015.06.073.
[5] Härmark L, Van Grootheest AC. Pharmacovigilance: Methods, recent developments and future perspectives. Eur J Clin Pharmacol 2008;64:743–52. https://doi.org/10.1007/s00228-008-0475-9.

[6] Clothier HJ, Crawford N, Russell MAMA, Buttery JP, Joanne Clothier H. Allergic adverse events following 2015 seasonal influenza vaccine, Victoria, Australia. Euro Surveill 2017;22:1–7. https://doi.org/10.2807/1560-7917.ES.2017.22.20.30535.
[7] Pal SN, Duncombe C, Falzon D, Olsson S. WHO strategy for collecting safety data in public health programmes: Complementing spontaneous reporting systems. Drug Saf 2013;36:75–81. https://doi.org/10.1007/s40264-012-0014-6.
[8] Shimabukuro TT, Nguyen M, Martin D, DeStefano F. Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS). Vaccine 2015;33:4398–405. https://doi.org/10.1016/j.vaccine.2015.07.035.
[9] After vaccination | The Australian Immunisation Handbook 2021. https://immunisationhandbook.health.gov.au/vaccination-procedures/after-vaccination (accessed February 15, 2021).
[10] Parrella A, Braunack-Mayer A, Gold M, Marshall H, Baghurst P. Healthcare providers' knowledge, experience and challenges of reporting adverse events following immunisation: a qualitative study. BMC Health Serv Res 2013;13:313. https://doi.org/10.1186/1472-6963-13-313.
[11] Mesfin Y, Cheng AC, Enticott J, Lawrie J, Buttery JP. Use of telephone helpline data for syndromic surveillance of adverse events following immunization in Australia: A retrospective study, 2009 to 2017. Vaccine 2020;38:5525–31.
[12] Conway M, Hu M, Chapman WW. Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. Yearb Med Inform 2019;28:208–17. https://doi.org/10.1055/s-0039-1677918.
[13] Paul MJ, Dredze M. Social Monitoring for Public Health. Synth Lect Inf Concepts, Retrieval, Serv 2017;9:1–183. https://doi.org/10.2200/S00791ED1V01Y201707ICR060.
[14] Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15:1–16. https://doi.org/10.2196/jmir.1933.
[15] McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. PLoS Negl Trop Dis 2017;11:1–15. https://doi.org/10.1371/journal.pntd.0005295.
[16] Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control 2015;43:563–71. https://doi.org/10.1016/j.ajic.2015.02.023.
[17] Delir Haghighi P, Kang Y-B, Buchbinder R, Burstein F, Whittle S. Investigating Subjective Experience and the Influence of Weather Among Individuals With Fibromyalgia: A Content Analysis of Twitter. JMIR Public Heal Surveill 2017;3:e4. https://doi.org/10.2196/publichealth.6344.
[18] Joshi A, Sparks R, McHugh J, Karimi S, Paris C, MacIntyre CR. Harnessing Tweets for Early Detection of an Acute Disease Event. Epidemiology 2020;31:90–7. https://doi.org/10.1097/EDE.0000000000001133.
[19] Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014;14:160–8. https://doi.org/10.1016/S1473-3099(13)70244-5.
[20] Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. Data Br 2017;10:122–31. https://doi.org/10.1016/j.dib.2016.11.056.
[21] Phan N, Bhole M, Ae Chun S, Geller J. Enabling real-Time drug abuse detection in tweets. Proc - Int Conf Data Eng 2017:1510–4. https://doi.org/10.1109/ICDE.2017.221.
[22] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015;54:202–12.

https://doi.org/10.1016/j.jbi.2015.02.004.
[23] World Health Organization. The Importance of Pharmacovigilance - Safety Monitoring of medicinal products. Who 2002:1–52. https://doi.org/10.1002/0470853093.
[24] Milstien JB, Batson A, Wertheimer AI. Vaccines and drugs: characteristics of their use to meet public health goals (English) 2015:1–40.
[25] Budhiraja S, Akinapelli R. Pharmacovigilance in vaccines. Indian J Pharmacol 2010;42:117.
[26] Almenoff J, Tonning JM, Gould AL, Szarfman A, Hauben M, Ouellet-Hellstrom R, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf 2005;28:981–1007. https://doi.org/10.2165/00002018-200528110-00002.
[27] Di Pasquale A, Bonanni P, Garçon N, Stanberry LR, El-Hodhod M, Tavares Da Silva F. Vaccine safety evaluation: Practical aspects in assessing benefits and risks. Vaccine 2016;34:6672–80. https://doi.org/10.1016/j.vaccine.2016.10.039.
[28] Clothier HJ, Lawrie J, Russell MA, Kelly H, Buttery JP. Early signal detection of adverse events following influenza vaccination using proportional reporting ratio, Victoria, Australia. PLoS One 2019;14:1–17. https://doi.org/10.1371/journal.pone.0224702.
[29] Sarker A, Gonzalez-Hernandez G. Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017. CEUR Workshop Proc 2017;1996:43–8.
[30] Weissenbacher D, Sarker A, Paul M, Gonzalez G. Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. Proc. 2018 EMNLP Work. SMM4H 3rd Soc. media Min. Heal. Appl. Work. Shar. task, 2018, p. 13–6.
[31] Weissenbacher D, Gonzalez-Hernandez G. Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task. 2019.
[32] Klein AZ, Alimova I, Flores I, Magge A, Miftahutdinov Z, Minard A-L, et al. Overview of the Fifth Social Media Mining for Health (SMM4H) Shared Tasks at COLING 2020. Proc. Fifth Soc. Media Min. Heal. Appl. Work. Shar. Task, 2020. https://doi.org/10.18653/v1/w19-3203.
[33] Lardon JJ, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: A scoping review. J Med Internet Res 2015;17:1–16. https://doi.org/10.2196/jmir.4304.
[34] Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. PLoS Comput Biol 2011;7. https://doi.org/10.1371/journal.pcbi.1002199.
[35] Larson HJ, Smith DMD, Paterson P, Cumming M, Eckersberger E, Freifeld CC, et al. Measuring vaccine confidence: Analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. Lancet Infect Dis 2013;13:606–13. https://doi.org/10.1016/S1473-3099(13)70108-7.
[36] Du J, Xu J, Song HY, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. BMC Med Inform Decis Mak 2017;17. https://doi.org/10.1186/s12911-017-0469-6.
[37] Chandrashekar PB, Magge A, Sarker A, Gonzalez G. Social media mining for identification and exploration of health-related information from pregnant women 2017;1. https://doi.org/arXiv:1702.02261.
[38] Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015;53:196–207. https://doi.org/10.1016/j.jbi.2014.11.002.
[39] Wang J, Zhao L, Ye Y. Semi-supervised Multi-instance Interpretable Models for Flu Shot Adverse Event Detection. Proc - 2018 IEEE Int Conf Big Data, Big Data 2018 2019:851–60. https://doi.org/10.1109/BigData.2018.8622434.
[40] Wang J, Zhao L. Multi-instance Domain Adaptation for Vaccine Adverse Event Detection 2018:97–106. https://doi.org/10.1145/3178876.3186051.
[41] Khademi Habibabadi S, Haghighi PD, Habibabadi SK, Haghighi PD, Khademi S, Haghighi

PD. Topic Modelling for Identification of Vaccine Reactions in Twitter. ACM Int Conf Proceeding Ser 2019:31. https://doi.org/10.1145/3290688.3290735.

[42] Zhai C, Massung S. Text Data Management and Analysis. 2016. https://doi.org/10.1145/2915031.

[43] Scikit-learn. scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation 2021. https://scikit-learn.org/stable/ (accessed May 24, 2021).

[44] sklearn.TfidfTransformer. sklearn.feature_extraction.text.TfidfTransformer — scikit-learn 0.24.2 documentation 2021. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html (accessed May 23, 2021).

[45] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 1st Int Conf Learn Represent ICLR 2013 - Work Track Proc 2013:1–12.

[46] Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. Proc Lr 2010 Work New Challenges 2010.

[47] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. 54th Annu Meet Assoc Comput Linguist ACL 2016 - Long Pap 2016;3:1715–25. https://doi.org/10.18653/v1/p16-1162.

Appendix A

As described in the Classification section, after some experimentation with the first set of imbalanced records, roughly balanced datasets were created, resulting in an initial (first phase) dataset of 3,519 tweets, and from the second phase of data collection a second dataset of 15,730 tweets, together with hold-out test datasets of 614 and 828. We trained classifiers on the first phase dataset of 3,519 tweets, but to re-evaluate classifiers with more data we combined both datasets and trained them on the resulting 19,249 tweets. Error: Reference source not found summarizes these numbers.

Table 6. Dataset Numbers

Stage	First Phase data collection	Second Phase data collection	Total
Into topic modelling	328,822	359,535	**688,357**
Minus filtered out by topic modelling	-310,021	-279,163	**-589,184**
After topic modelling	**18,801**	**80,372**	**99,173**
Minus data preparation and balancing	-14,668	-63,814	**-78,482**
For classification training	**4,133**	**16,558**	**20,691**
For training and validation	3,519	15,730	**19,249**
For testing	614	828	**1,442**