MONASH University

Vaccine safety surveillance using social media data

Sedigheh Khademi Habibabadi

A thesis submitted for the degree of *Doctor of Philosophy* at
Monash University in *2020*
Faculty of Information Technology

Main Supervisor: Dr. Pari Delir Haghighi
Associate Supervisor: Professor Frada Burstein
Associate Supervisor: Professor Jim Buttery

Copyright notice

Abstract

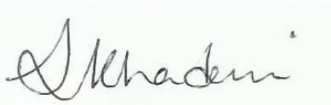
This dissertation is being published during an extraordinary time, under the shadow of the worldwide spread of COVID-19, with the world hoping for vaccines that would allow us to breath freely again, to give us back our freedom to move, to emerge from lockdowns, fear and frustration. Although vaccine manufacturers are being careful to put vaccines through well-established testing processes, there are new vaccine technologies on trial and there is an intense public health and political pressure to bring them to market. It is inevitable that as vaccines are disseminated to the general population there will be reports of adverse health-related events in relation to vaccine distributions, some of which will be genuine vaccine reactions, either expected or untoward. It is vital that reports of adverse events are continually monitored to help ensure a rapid response to any emerging issues with vaccine safety and effectiveness.

Traditional monitoring for Adverse Events Following Immunisation (AEFI) relies on various established reporting systems, where there is inevitably a lag between an adverse event following a vaccine and the reporting of it, and subsequent processing of reports. Therefore, it is desirable to try and detect AEFI earlier, ideally close to real-time, and monitoring social media data holds promise as a resource for this. However, social media users relaying experiences of adverse events following vaccination are difficult to detect – there is an overwhelming amount of other vaccine and virus-related conversations swamping social media platforms. This research is dedicated to proving that useful Vaccine Adverse Event Mentions (VAEM) *can* be detected in social media, using Twitter as a data source, and applying natural language processing techniques to successively filter out unwanted messages to bring VAEM to light.

This research has developed a VAEM-Mine method that combines two stages of topic modelling with classification to extract around 90% of all VAEM posts from a Twitter stream, with a high degree of confidence. This is a significant achievement, as VAEM posts constitute less than 2% of all vaccine-related Twitter posts. The research also presents a taxonomy of vaccine-related Twitter posts, datasets of VAEM Twitter posts and detailed reporting on the most effective approaches to topic modelling and to classification of extracted posts, in relation to varying data volumes. This work provides a methodological foundation for potential near-real time monitoring of social media VAEM to augment existing signal detection systems, maximising the ability to detect unsafe vaccines rapidly.

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: 

Print Name: Sedigheh Khademi Habibabadi

Date: 20/10/2020

Acknowledgements

I would like to express my deepest gratitude to my supervisors. To Dr. Pari DelirHaghighi, for her guidance in structuring the research and composition of the dissertation, her attention to detail, and her unending patience throughout the refinement of the writing. Her observations have provided a clear light for my research focus.

To Professor Frada Burstein for being my mentor throughout many stages of my life and educational journey at Monash University. She has continuously nurtured, encouraged, and supported me in my goals and has been delighted in my personal and professional progress. I wholeheartedly believe that my academic success is due in a large part to her support and the inspiration I have gained by knowing her.

To Professor Jim Buttery for providing the research motivation, and for being so positive and practically active whenever I have needed him. He above all has championed this research because he knows what is at stake, that improving our ability to detect potential vaccine safety signals early is not merely an academic concern but can save lives. I am profoundly fortunate that he has created opportunities to put this research into practice, the years ahead are already opening creative and professional pathways because of Jim's influence.

I thank my dear husband and best friend, Christopher Palmer. This project could not have evolved without his ongoing love and support. From taking care of family, to guiding me in technical aspects of my work, to proof reading and editing – he is always enthusiastic, perceptive, and my deeply supportive and loving partner in my life and work.

I started this journey with my infant Noora and three years old Zoha and I would like to thank my darling daughters for their patience when I took time away from them to study, but I thank them for so much more! They have become the whole meaning of our lives; and our hearts beat in time with the rhythms of their beautiful souls.

To my sister Mahsa who throughout my life has been my most loyal and dearest friend and ally. She has extremely generously given herself to being a pivotal part of our family life, taking care of our family while I needed to take time to study. She completely exemplifies the depth and meaning of the love of a family. I am thankful to my parents for their unconditional love, support, and dedication to our academic and life success.

Many thanks also to the wonderful members of Graduate Research team at Monash Faculty of Information Technology, particularly Ms. Helen Cridland and Ms. Allison Mitchell for being always understanding, cheerfully supportive and accommodating.

I also extend my gratitude to the team at SAEFVIC, especially Dr. Jock Lawrie for his help in reconciling my Victorian dataset with their data, and Dr. Hazel Clotheir for her valuable input. I want to thank Dr. Lan Du for taking time to read my work and for his suggestions to improve it, and Dr. Reza Haffari for his guidance in the early part of my study. I am fortunate to be engaged in natural language processing research where some of the most exciting developments in artificial intelligence are rapidly advancing, driven by a collaborative and generous community of great thinkers and teachers. It is a great time to be learning.

He brings them forth from the shadows into the light

Quran 2.257

To my dear daughters
Zoha and Noora,
May you always find
your way in life
and fully realize
the immense love
that guides you

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADR	Adverse Drug Reactions
AEFI	Adverse Event Following Immunization
AEFI-CAN	Adverse Events Following Immunisation – Clinical Assessment Network
BERT	Bidirectional Encoder Representations
BiGRU	Bi-directional Gated Recurrent Unit
BiLSTM	Bi-directional Long Short-Term Memory
CBOW	Continuous Bag-of-Words
CCV	CalibratedClassifierCV
CNN	Convolutional Neural Networks
CNN	Convoluted Neural Networks
COVID-19	Coronavirus Disease of 2019
DMM	Dirichlet Multinomial Mixture
EHRs	Electronic Healthcare Records
GRU	Gated Recurrent Unit
LDA	Latent Dirichlet Allocation
LRCV	Logistic Regression CV
LSTM	Long Short-Term Memory
ML	Machine learning
MMR	Measles, Mumps and Rubella
MMR	“Match Making Rating” in a gaming context
NLP	Natural Language Processing
RNN	Recurring Neural Network
RoBERTa	Robustly Optimized BERT Pretraining Approach
SGD	Stochastic Gradient Descent
SLDA	Supervised Latent Dirichlet Allocation
SRS	Spontaneous Reporting Systems
TGA	Therapeutic Goods Administration
VAEM	Vaccine Adverse Event Mention
VSD	Vaccine Safety Datalink
WHO	World Health Organization

1 Introduction

Vaccinations are one of the main components of public health programs and have significantly contributed to reducing mortality and morbidity rates of communicable diseases. At the time of completing this thesis the importance of vaccines is highlighted more than ever, as the world combats the COVID-19 pandemic. Vaccine uptake must be as complete as possible for effective disease prevention, with vaccinations of vulnerable children being a key feature of a vaccinated population. The safety of vaccines is vital, both for vaccine recipients and for public trust and confidence in vaccine programs (Jacobson et al., 2001). Monitoring for vaccine reactions is a key component of ensuring vaccine safety.

Social media is increasingly becoming an additional source for health-related information, including disease, drug and vaccine-related reaction mentions. 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 (Paul & Dredze, 2017). 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.

This thesis seeks to establish the usefulness of social media for vaccine safety monitoring by applying current natural language and machine learning technologies to the task of detecting and extracting vaccine adverse reaction mentions from the enormous volume of other vaccine-related social media discourse.

1.1 Vaccines and vaccine safety

Continued acceptance of vaccination programs is vital. WHO defines vaccine hesitancy as " the reluctance or refusal to vaccinate despite the availability of vaccines" and declares it as one of ten threats to global health (WHO, 2019). When most of the population have immunity against a contagious disease then the chance for an outbreak of the disease is minimized or even eliminated due to "herd immunity". Vaccines are developed to help us to achieve immunity without having to go through the suffering and attrition of developing natural immunity through the ravages of an epidemic or pandemic. Therefore, public confidence in the safety and effectiveness of vaccines is necessary for general uptake, though additionally some form of mandatory vaccination is often considered (Hardt et al., 2013). Mandatory vaccination

programs are broadly defined as vaccinations that every child must receive by law and can vary from soft/flexible to hard/inflexible (MacDonald et al., 2018). An example of inflexible law is the "No jab no pay" national and "No jab-no play" state legislation in Australia (Abbott & Morrison, 2015; National centre of Immunization, 2017), where financial penalties or social restrictions pertain for non-compliance.

Public concerns over vaccine safety are ironically highlighted because of the great success of vaccines, the dreadful diseases of the past that are prevented by vaccines are no longer generally experienced by the population. Except in countries still battling epidemics, and until the experience of the current coronavirus pandemic for everyone else, an appreciation of what vaccines are protecting us against has waned. Consequently, people have focused their attention on vaccines themselves, and especially on their safety (Larson et al., 2011). Media reports of reactions or side-effects following vaccinations can quickly lower trust in a vaccine and ultimately affect its acceptance and uptake, particularly when these reports concern children and are sensationalized. If the vaccine in question is for a major disease, then lower uptake could result in a resurgence of the disease (Lantos et al., 2010); or in the case of a new disease like COVID-19, fail to achieve vaccine-enabled herd immunity.

Vaccine authorities name an untoward medical effect after receiving a vaccine as an Adverse Event Following Immunization (AEFI), or more simply as an "adverse event" — see Section 2.2.1 for a full definition. There is no implicit causality or required level of severity in an AEFI, it is essentially to take note of *anything* relating to an individual's health in the context of a recent immunization. Noting and investigating AEFI is a key component in monitoring vaccine safety.

To ensure that vaccines are safe and effective, it is of utmost importance that vaccines are manufactured and produced safely, and that they are thoroughly tested and proved to be safe and efficacious before distribution. To this end, a pre-licensure process takes place where vaccines go through three trial phases, where commonly encountered and mostly mild and temporary adverse events are documented, and vaccine efficacy is established (Salmon & Halsey, 2016). Due to the adherence to high safety standards, serious adverse events are rare and so not generally encountered until a vaccination has been implemented within a population generally (Council for International Organizations of Medical Sciences & WHO, 2012).

Distribution and application of vaccines to target populations requires a number of ongoing safety-oriented measures, starting with correct transport and storage (usually refrigerated), and proper administration with sterile equipment — failures in these areas have resulted in great harm (Evans et al., 2016). Additionally, there needs to be ongoing monitoring for untoward

side-effects from a vaccine, particularly for a pattern that may indicate a previously unknown issue with a vaccine, or issues with local handling and administration of a vaccine.

Most detection of vaccine problems is based on noting an increased incidence of reporting of an adverse event compared with the "normal background rate" of the event, the increased incidence is described as a "vaccine safety signal". Therefore, timely data collection and analysis are essential for detecting any issues that may arise after a vaccine has been administered generally, and it is vital to look at these vaccine safety signals from as many sources as possible (Crawford et al., 2014).

1.2 Social media monitoring

Traditional reporting is seldom patient-driven, the average person does not actively seek to communicate information about a health issue via an established reporting system, instead they go to their doctor for advice and help, and it is the doctor's choice to report any issues that are notifiable or they feel are noteworthy, which requires additional actions by the doctor (O'Shea, 2017). However, a visit to a health professional is not always possible, or even if possible is not necessarily the first point of contact, people these days are inclined to consult with the internet, looking for and sharing information outside of any traditional reporting systems (Gualtieri, 2009; Tan & Goonawardene, 2017). Data resulting from these online activities is labelled "user generated" in the health data domain, and is a potentially valuable source for health studies, and increasingly a component of surveillance systems.

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, although not something a health system can respond to individually, 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.

Natural Language Processing (NLP) techniques are essential in social media monitoring, as the enormous amounts of data that are captured can only be handled effectively by computers. NLP is the use of computing-based technologies to understand language and is a very active area of artificial intelligence research. NLP encompasses a range of *machine learning* tools, including topic modelling to discover the semantic meaning or topics of texts, and increasingly uses powerful neural network-based technology to understand, translate, and generate text and speech — an approach known as *deep learning*. Recent years have seen major advances in these technologies, so that every year a new state-of-the-art improvement is announced and

made available by the research laboratories of major organizations and universities like Google, Facebook, OpenAI, Stanford, and the Allen Institute for Artificial Intelligence.

1.3 Vaccine adverse event mentions

Vaccines belong to the broad category of medicine, but in a subcategory known as "biologicals" (Milstien et al., 2015). Unlike drugs 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 of time, to enhance their immune systems' ability to combat a pathogen. In contrast to those who are taking a drug to help to cure a disease or to treat unwanted symptoms, most people receiving a vaccine are not ill. Therefore, there is a deferred individual benefit to taking a vaccine, and consequently a very low acceptance of risk regarding vaccines (Budhiraja & Akinapelli, 2010). Additionally, the pathophysiology of vaccine adverse events is not as well defined as those of adverse drug reactions - due to vaccine's complex biological nature and interaction with the immune system, a reaction caused by a vaccine could be caused by any of its multiple ingredients or even an error in administration (Almenoff et al., 2005). Added to these issues, as vaccines are administered to such a varied population, reactions may be exacerbated by unknown underlying conditions in vaccine recipients, compared with the specific groups targeted for traditional drug therapies. Furthermore, a vaccine's "time to market" may be curtailed such as has occurred in the COVID-19 pandemic, and not provide 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, and 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 sever sequelae (such as increased rates of fever may be a marker for increased febrile seizures (Mesfin et al., 2020)), and may also play a major role in affecting vaccine confidence (Di Pasquale et al., 2016). 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 pharmacovigilance is considered to be different from pharmacovigilance of other medications and has specialized regulatory guidelines and surveillance systems around the world; examples include the US Vaccine Adverse Event Reporting System (VAERS) (Chen et al., 1994) in the United States and Surveillance of Adverse Events In the Community (SAEFVIC) (Clothier et al., 2011) in Australia. Vaccine 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. (2019): "*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*"'.

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 formal AEFI reporting, and what are typically thought of as being severe adverse vaccine events, or drug reactions for that matter. Which is to say, VAEM are conversations, ideally gathered in volume, that contain information that might be those common AEFI that are so elusive to traditional reporting.

1.4 Problem Statement

Spontaneous passive reporting systems, where predominantly health professionals log data about adverse events they have observed, have been in place in many countries for decades — but even so, significant incidents may not be reported on in a timely manner (Armstrong et al., 2011). For instance there were serious delays in reporting reactions to the TGA during the 2010 flu vaccine in Australia, a subsequent ministerial review criticised the existing processes as contributing to delays in cancelling the vaccine program, and made recommendations about improving them (Stokes, 2010). Spontaneous reporting systems often suffer from underreporting and lag between the occurrence of the issues and the time it becomes known by the authorities (Isaacs et al., 2005). The Australian Immunization Handbook states that vaccine providers should use their clinical judgment when deciding to report an event (*After vaccination | The Australian Immunisation Handbook*, 2021). Consequently, the existing reporting system is potentially filtering out AEFI that individual practitioners decide not to report. Parrella et al. (2013) studied Australian healthcare providers' knowledge and the challenges of AEFI reporting concluded 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. For instance, the study found that Paediatric Emergency Department consultants, overall, would only report severe, "life-threatening" events. Apart from potential underreporting by health professionals, Mesfin et al. (2020) point out that reporting on AEFI that occur *after* a patient has gone home depends on patients or their caregivers returning to the clinic or visiting an emergency department or hospital. Without such

a visit, less severe AEFI are unlikely to be captured. Their study suggests that "AEFI-related calls" from telephone-based triage systems, such as the Nurse-On-Call (NOC) system, offers opportunities for additional near real–time syndromic surveillance of AEFI, with the potential to identify sever AEFI signals earlier.

In conclusion, it is desirable to try and detect a wider range of adverse events, and ideally close to real-time. This research seeks to establish if social media can usefully contribute to early detection of a broader range of vaccine adverse events, by first confirming that social media posts can contain textual information that is interpretable as AEFI, and what effective techniques can be used to isolate such posts.

Extensive use of social media has provided a platform for sharing and seeking health-related information. Social media monitoring of health-related conversations offers a separate source of information that can be used to supplement and corroborate health-related signals coming from established health reporting systems. Real-time monitoring of social media discussions about health-related issues opens the possibility of early warning detection of emerging health crises (Steele, 2011).

However, there are numerous challenges in filtering and interpreting the immense volume of social media data for true signals of current health issues, and for this study, vaccine-related adverse event mentions. The challenges encountered when filtering largely stem from the disparity between the huge volume of social media data and difficulties in extracting relevant information from it. Difficulties encountered in interpretation are due to the informal language and structure of social media posts.

Even within the context of formally established health-related reporting systems (Spontaneous Reporting Systems - SRS), the characteristics of self-reporting about health problems by lay people are quite different from reporting by health professionals, and a topic of ongoing analysis for reliable information (Krska et al., 2011). Self-reporting in the context of social media conversations is by contrast a completely unstructured information source and it is not clear yet how it can best be utilized.

In the health domain professionals use formal reporting methods and a more agreed upon vocabulary, and data is often in a structured form. Where clinical notes exist as free text, then there are difficulties in extracting codified meanings from them, and this is a current research problem (Yadav et al., 2018), but the text usually exists in a reliable database and often contains common terms that are amenable to codification using rule and machine-learning based natural language processing (Khademi et al., 2015). To reiterate, social media conversations on the

other hand are extremely widespread and wide ranging, so there are challenges in both finding potentially relevant data and in interpreting it to see if it is useful.

There are various challenges involved in finding potentially relevant data in social media posts:

- There are many different forums for people to discuss their concerns, and these in turn are based on different technologies (e.g., Twitter, Facebook, Reddit). That is, data is not readily accessible from just a few well understood sources, as it is with a known system like a health-care database.
- Social media forums evolve, new platforms emerge, favourites appear and disappear — these in turn are centred around different demographics. For instance, Snapchat and Instagram may be favoured by teenagers but Facebook might be preferred by their mothers. Geographical variations also exist, for instance WeChat and Qzone are Chinese-specific messaging and social media applications. Therefore, social media surveillance strategies must be continuously adjusted to account for a dynamic social media landscape.
- Terms that may have a specific meaning in a health-care context may be used in other contexts with a completely different meaning or application, and these can provide false signals. For example, from the vaccine-related field the term MMR stands for "Measles, Mumps and Rubella" and used to refer to the triple combination vaccine for these diseases. In a health-care database it can be safely assumed that the use of MMR is regarding the vaccine, but on-line gaming communities use MMR as an acronym for "Match Making Rating" and MMR is possibly used online more often in the gaming context. The words virus and viral have completely different applications outside of health-care, *virus* is used to describe malicious software and *viral* can refer to marketing techniques or the popularity of a rapidly emerging internet-based meme.
- Beside the problem of the polysemous nature of recognised terms, people discussing their health issues are not necessarily going to use them. For instance, a mother may refer to her daughter's reaction to a first MMR vaccine as "DD really sick and rashy with her 1 YO jab". However, some of these terms have many other meanings: "DD" is commonly used as here for "darling daughter", but can also be used for "direct download", "direct debit", "disk drive" etc. "Sick" can be an affirmation that something is really great, and a "rashy" is a type of summer protective clothing! Besides "year old", YO can mean "your" or "hi"; and "jab" is used in countless contexts.

- There are no dictionaries containing all the possible variation of medical concepts. Therefore, simply searching for the colloquial equivalents of terms of interest is likely to lead to an overload of bogus noise; however, searching for terms that are only used in a medical context is likely to somewhat limit conversations to those coming from health-professionals, which for the study of emerging trends is not a broad enough search scope. Non-technical terms used to discuss health issues are described as being "consumer expressions" in contrast to the recognized professional terms use by health professionals (Farzindar & Inkpen, 2015), and they seldom find their way into dictionaries and corpuses of recognized medical terms that would normally be used as the basis for pattern matching with medical notes.
- Traditional NLP tools such as part-of-speech taggers, parsers and lexicons may be insufficient to successfully classify social media conversations (Sarker et al., 2015). A traditional approach may be too complex to formulate and maintain, due to the variation in the language of social media. On the other hand, using automated approaches require large volume of examples containing the variations of language used.
- Finding relevant health signals from social media text is more complicated than using other sources of data such as healthcare or medical data. There are no formal structures in the texts where a mention of a specific vaccine reaction is interpretable as a potential safety signal, instead there are numerous wildly different discussions and articles about vaccines in social media. Many of these might be mined for sentiment and opinions, but for texts to be considered as potentially useful for safety signal detection they should either indicate that a recent vaccine may be seriously affecting an individual's health; or more likely, that there is an increased incidence of reporting of expected reaction-mentions in relation to a vaccine program.
- When machine learning approaches are used to classify data there usually needs to be a corpus of labelled data for the learning process. Labelling is a challenge when dealing with social media conversations: language is highly varied and useful text may be buried in irrelevant surrounding conversation (Cocos et al., 2017). Labelling may require preparatory tasks which themselves involve language processing.

In summary, vaccine safety surveillance is enhanced by responsive reporting, and there is a possibility of obtaining near real-time vaccine adverse event mentions through social media streaming data, provided the numerous challenges in obtaining clear adverse event mentions can be overcome.

1.5 Research aims and objectives

The inspiration and aim for this study are to contribute to research on vaccine safety surveillance. This was motivated by a leading Australian vaccine expert's request to investigate whether social media streams might contain useful vaccine adverse event mentions, that could potentially contribute to earlier detection of emerging issues with vaccines, that might constitute a "safety signal". The importance of this is described earlier in the problem statement: the earlier that safety signals can be detected then the greater the opportunity to act to prevent harm. The remit was to thoroughly explore, understand and document the problem and solutions of how to obtain vaccine adverse event mentions. Ideally, solutions should apply established techniques that would be accessible and utilizable by vaccine safety monitoring authorities such as the Therapeutic Goods Administration (TGA) and the Adverse Events Following Immunisation – Clinical Assessment Network (AEFI-CAN).

The key objective of the study is to develop and describe an understanding of how to differentiate vaccine adverse events mentions from the myriad of other conversations that mention vaccinations or personal health issues. It was theorised by the author that the kind of language used by people describing personally experienced vaccine-related health issues would be distinguishable from people discussing vaccines in general, from anti/pro-vaccine arguments, from news articles about vaccination, and from other personal health mentions and of course, unrelated texts. The research set out to empirically verify this supposition using existing modern natural language processing techniques, aiming to combine commonly understood practices to create an effective method for detecting vaccine adverse event mentions from social media. The research also intended to thoroughly explore and document the problem and solution domain.

The following research questions were proposed and addressed in this doctoral project:

1.5.1 Research questions

The aim of the research was to determine if social media surveillance could assist with detection of vaccine safety signals. We sought to answer the following questions:

RQ1 — What effective techniques can be utilized for identifying posts containing vaccine adverse event mentions (VAEM)?

RQ2 — How can a comprehensive dataset be assembled and labelled which will enable both this research and further research into vaccine discourse in social media?

RQ3 — What taxonomy of vaccine-related Twitter posts can be derived to assist with analysis of Twitter conversations regarding vaccines?

Relevant posts (VAEM) are differentiated from commentary and voicing of opinions relating to vaccines and are describing personal or familial health events following vaccination.

1.5.2 Research design

This research proposed, developed, and validated approaches to answer these questions, following a prescribed but dynamically evolving research design:

1. The design firstly consisted of exploring and defining the problem and solution domain by reviewing the literature around vaccine safety surveillance, the challenges in using social media in public health surveillance, and social media mining techniques. This flowed into experimental exploration of and training in the various technologies, especially in machine learning and deep learning — enough to apply them practically to solving the problems of the research. Where needed a domain expert and machine learning and NLP experts were consulted for their theoretical and practical help. These steps were foundational for answering the research questions.
2. Vaccine-related Twitter data was collected, applying insights obtained from the literature review; through studying the use of the Twitter API; and from data exploration, including the assessment of data published by research laboratories and from pre-existing social media data.
3. Data collection and experiments on the data were conducted in two phases. This enabled evaluation of the techniques. The first phase of data collection gathered 6 months of data and was used to develop a topic modelling approach, the second phase added a further 6 months of data, which was used to evaluate the trained topic models when applied to a new dataset. Classifiers were able to be assessed with 6 months of data and reassessed with a year's amount of data, thus allowing their performance to be measured relative to data quantity. These measurements were essential for understanding the effectiveness of the techniques and data importance.
4. Topic models were assessed to understand the semantic themes of the data. This led to the identification of texts that include vaccine adverse event mentions, but also a realization that the intrinsic scoring techniques of the various models were insufficient for the task of identifying the configuration needed for the ideal of gathering VAEM into one topic, or for comparing models. A scoring system based on labelling a small number of records was developed and proved to be highly successful for tracking when

VAEM-containing texts are focused into a topic. A vaccine-related taxonomy was developed using the topics suggested by the topic models. Data from the best performing topic model was extracted and labelled for classification.

5. An extensive range of classifiers were trained and compared on the labelled data, and this was performed twice — first on data collected over 6 months, then expanded to a larger dataset collecting for a year. This approach enabled the research to measure the suitability of classifiers in relation to data size, which turned out to be a significant effect. The models assessed include the most powerful deep learning models at the time, which use a "Transformer" architecture.
6. A method, named VAEM-Mine, was designed for extracting vaccine adverse event mentions from Twitter data — it encapsulated the workflow and techniques required to combine the most effective of the processes for extracting vaccine adverse event mentions from Twitter data. The method is applicable to any similar requirement and data.
7. All these steps were thoroughly documented, with the goal of sharing the insights gained during the research and to allow for reproducible research, especially regarding how to practically mine vaccine adverse event mentions from social media.

1.6 Research contribution

The motivation and aim of the research were to contribute to vaccine safety surveillance, by research results which focus on the role of social media in earlier detection of potential vaccine safety signals. The research first needed to answer whether vaccine adverse event mentions can be reliably detected in social media streams, and if so then to deliver information about effective techniques for vaccine adverse event mention detection, accompanied by a vaccine-related dataset and a vaccine taxonomy. This study confirmed that social media *can* be used as a source for vaccine safety surveillance, by proving that Twitter posts can be effectively mined for vaccine adverse event mentions.

This research can be described as the confluence of the needs of Immunisation Research and Surveillance (IRS) organizations, and the available infrastructure and knowledgebase. The research adds to prescriptive knowledge as it thoroughly explores and describes the challenges and solutions involved in detecting vaccine adverse event mentions in social media. Further, the research contributes a highly effective method for identifying vaccine adverse event mentions from the vast majority of other vaccine-related posts. The proposed VAEM-Mine method is comprehensive, efficient, easily implementable, and generally applicable to any

similar problem of identifying personal health mentions based on the types of language used in them.

The contributions made by this research are:

1. A very positive affirmation that social media surveillance can assist with vaccine safety signal detection. The research finds that combinations of readily available NLP tools can be used to extract almost all vaccine adverse event mentions from Twitter data, while eliminating irrelevant posts. The extracted data has been confirmed as fit for the purpose of safety signal detection by an expert's analysis and by a comparative trend study. 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, to help address the deficiencies of timeliness and under-reporting of passive reporting systems.
2. The research provides explanations of the use of the NLP technologies that have been most successful in isolating vaccine adverse event mentions, to answer the first research question. Crucially, this includes describing a highly effective and easily implementable method for identifying the best performing topic models for the specific task of identifying almost all relevant VAEM posts, and the subsequent use of the current best deep learning models to then refine the extracted data.
3. The topic modelling phase excluded many vaccine-related tweets and produced data that was either VAEM or like VAEM (personal health mentions and discussions), which collectively can be characterized as "VAEM-like". These were labelled for classification, resulting in a *balanced* labelled dataset of 20,777 tweets and a larger labelled but *imbalanced* dataset of 83,891 tweets. These datasets can be shared under Twitter licensing conditions and satisfy the second research question regarding a comprehensive dataset.
4. A taxonomy of vaccine-related Twitter posts, and so answering the third research question. The taxonomy will facilitate understanding the type of vaccine-related discussions on Twitter. The social media posts investigated by the research are evaluated against the taxonomy to determine what kinds of posts predominate. This will aid understanding of the usefulness of social media for research into a range of vaccine-related subjects.
5. In technical terms, the research describes:

- 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.
- A significant capability by the VAEM-Mine method 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 was able to identify VAEM with an F1-Score of 0.91.
- Detailed reporting and comparisons on a range of classification models, from standard machine learning models, to custom rule-based approaches, through to 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.

Additionally, the research offers observations into the data collected and its potential utility. It was found that Twitter posts regarding potential reactions following vaccination, although incredibly varied in language, were consistent in nature, being personal complaints of usually common side-effects following a recent vaccination. These were confirmed as the expected data by the domain expert. It became clear that the usefulness of social media for detecting potential vaccine safety signals depends on being able to reliably collect most of these posts, so that trends can be detected in changing volumes of vaccine adverse event mentions. Specific descriptions of unusual adverse events were vanishingly small and a targeted detection of them

was not considered to be a useful direction to take. However, it should be noted that by reliably collecting most of the AEFI-related social media posts, detection of rare rapid-onset events can be also considered. Evaluation of the usefulness of vaccine adverse event mentions for early detection of vaccine safety signals could be further performed by vaccine authorities — by examining the data; by testing and refining the method; and by using the method (or methods like it) to collect social media posts over an extended period, or for a specific purpose.

1.7 Structure of the thesis

The thesis is structured as follows:

Chapter 1 – *Introduction* — Introduces the context of the research: the need for vaccines and the requirement for vaccine safety surveillance, and to exploit social media monitoring to assist with that surveillance. Includes the problem statement, research aims and objectives including the research questions and research design, which includes the need to determine effective techniques for detection of Vaccine Adverse Event Mentions (VAEM). Summarizes the research contributions. Finishes with a summary of the thesis structure.

Chapter 2 – *Background and Related Literature* — Provides the background and the related literature around vaccine safety and its surveillance, the use of social media in public health monitoring and various automated Natural Language Processing (NLP) methods employed in the health-related surveillance domain.

Chapter 3 – *Research Design* — Describes the research approach used to understand and analyse the problem and solution domain of social media for vaccine safety surveillance, then the NLP techniques that will be used to create a method for detecting VAEM.

Chapter 4 – *Data collection and preparation* — Explains the data gathering process and various pre-processing techniques used to get the data ready to process via topic modelling. Then discusses the various datasets constructed over the two phases of data collection, starting with datasets that were processed through topic modelling, then the subsequent datasets that were used with classification.

Chapter 5 – *Topic modelling* — Introduces topic modelling, including topic modelling scoring approaches. Describes the data preparation specific to topic modelling: lemmatization and vectorization etc. Discusses the two-stage topic modelling that applies a customized

scoring technique on a limited number of labelled records, to effectively identify and isolate vaccine adverse event mentions from a large amount of other vaccine-related Twitter posts. Shows the taxonomy that was derived from the topic models. Includes an evaluation of the effectiveness of the topic model scoring approach.

Chapter 6 – *Classification* — illustrates the classification methods, including standard classifiers, a rule-based classifier, and various deep learning models, which were employed to increase the precision of isolating vaccine adverse event mention-containing posts. Two phases of data collection and the effects of data size on different classifiers' performance are detailed, and the model scores are compared.

Chapter 7 – *Evaluation* — presents various evaluations of the use of topic modelling and classification. Figures are used extensively to aid understanding of the research results.

Chapter 8 – *Discussion and Conclusion* — Summarizes the research, including a more detailed description of how the research addressed the research questions. Describes the *VAEM-Mine Method*, which formalises the various processes that were used in this research as a method, which can be used to reproduce the research approach to similar problems. Explains the contribution, discusses the limitations and future research.

2 Literature Review

2.1 Chapter overview

Detection of vaccine safety signals depends on various established reporting systems, where there is inevitably a lag between an adverse reaction to a vaccine and subsequent reporting of it. With the advent of the internet, user-generated information on the web has become another source for vaccine-related data. This research endeavours to explore how social media surveillance can assist with detection of adverse vaccine reactions.

This background chapter is made up of four sections:

Vaccine safety — an explanation of the importance of vaccine safety; the stages of vaccine safety surveillance and how vaccine surveillance differs from monitoring of other medicines. The section concludes with an evaluation of the suitability of user-generated data, such as found in social media.

Social media data sources for public health studies

For public health studies, there are two major categories of social media platforms. General-purpose domains such as Twitter, Facebook and Reddit; and domain-specific platforms such as PatientsLikeMe, DailyStrength and drugs-forum. The domain-specific social media forums provide data that focuses on specific health issues and interests. In contrast, general-purpose social media conversations cover a broad range of topics and are more suitable for discovering current trends and common subjects. Topics like the COVID-19 pandemic and vaccines in general are discussed by everyone and are well mentioned in general-purpose social media.

Twitter, Facebook and Reddit are the most used social media sources for research in public health (Dol et al., 2019; Singh et al., 2020; Tang et al., 2018). The privacy and data sharing policy of these platforms impacts their popularity in research. Facebook, although the largest social network is very seldom used in research, mainly due to its strict privacy policies. Reddit has been a source of data for public health studies — however, it is used a lot less than Twitter. Compared to Reddit, the frequency of information dissemination in Twitter is much higher, the interval between posts is measured in minutes rather than hours. Additionally, the short text requirements of tweets encourages focused conversations that use more prevalent terms (D. Choi et al., 2016). According to Priya et al. (2019), tweets reduce the articulation of biased and extreme views, compared to Reddit posts. They also find that Twitter conversations can gather a greater momentum than Reddit posts, which can result in a news event being sustained

for much longer in Twitter, and that the spread, availability, and low inter-arrival times in Twitter posts make them very suitable for real-time monitoring, including emergencies.

Amongst the general-purpose social media platforms Twitter is most widely used as an additional source of data in the pharmacovigilance domain and in public health research (Edo-Osagie et al., 2020; Gupta & Katarya, 2020; Singh et al., 2020). Its real-time nature, high volume of messages and public availability makes it a suitable alternate source of data for surveillance of public health (Lardon et al., 2018).

Surveillance using social media — a review of the use of social media in health-related surveillance, particularly for disease surveillance, medication safety and personal-health mentions, and social media posts relating to vaccines and vaccinations.

Social media data processing — describes data collection and text pre-processing, including the tokenization and vectorization of text data to prepare it for machine learning.

Machine learning methods in text — explanations of text classification and content analysis; supervised and unsupervised learning, topic modelling, deep learning and transfer learning.

Each subsection finishes with a review and analysis of its main points, and the chapter concludes with a synthesis of the aspects of the literature review that are important for understanding the direction of this research.

2.2 Vaccine safety

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 (Chen, 1999). As high levels of immunization are achieved, the risk of contracting the diseases being vaccinated against diminishes, but the relatively small risks associated with vaccinations assume greater importance. A divergence emerges between individual and community risks vs benefits, and individuals only see the risks they are familiar with, those associated with vaccines (Salmon & Omer, 2006). Consequently, vaccine uptake may diminish, which in turn means that disease outbreaks may occur, and attention is again paid to the benefits of vaccination. This dynamic of initial vaccine uptake with accompanying disease suppression, followed by a loss of vaccine confidence and its ramifications, has been explained diagrammatically by Chen (1999) as depicted in Figure 1.

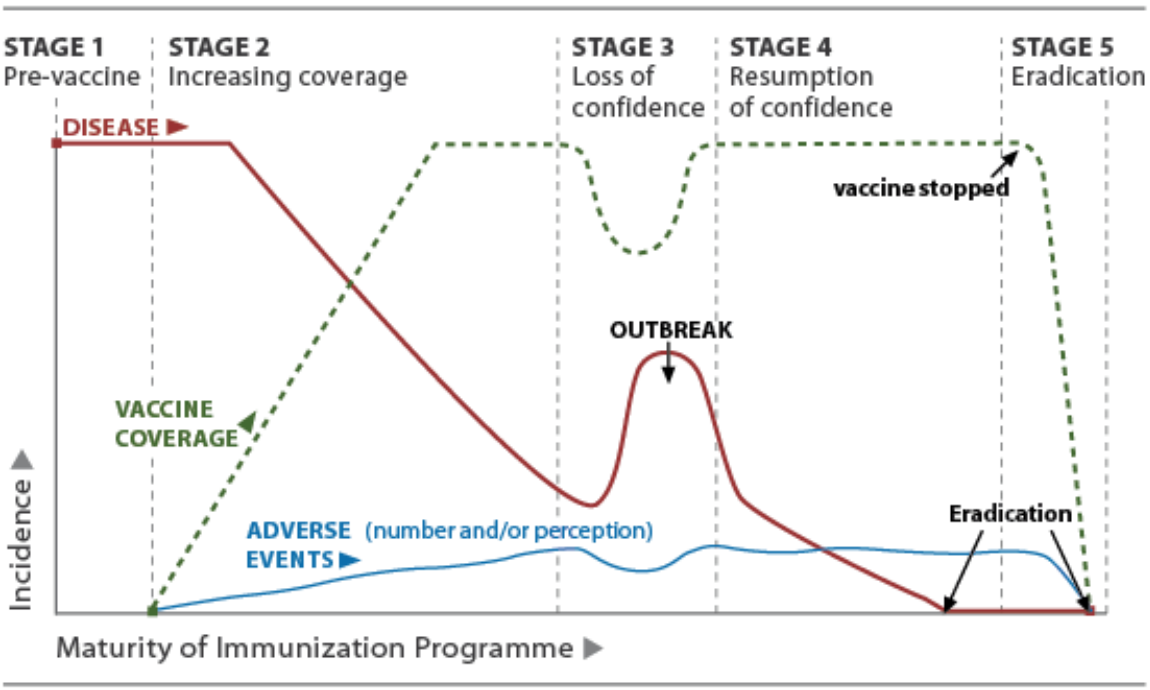


Figure 1: Potential stages in the evolution of an immunization program

In addition to the growing perception of risks associated with traditional vaccines, new vaccines are being constantly developed for diseases which were not previously considered, such as annual influenza, rotavirus, hepatitis, meningococcal meningitis, and human papillomavirus. As a result, vaccines and vaccine schedules are becoming much more complex; consequently, the overall risk of vaccine-related reactions and illnesses increases, leading to a higher perception of risks (*Vaccine history timeline*, 2018).

A greater awareness of and lower tolerance for the risks associated with vaccinations also accompanies changes in parenting approaches, where ideas about natural living, individual freedom and informed decision-making mean that parents are much more critical of anything that might risk their children's health, and more willing to resist societal pressure (and legislation) to vaccinate (Freed et al., 2010).

There is a growing movement of those who are vaccine hesitant as perceptions of vaccine risks grow, and ideas such as the corruption of "big pharma" and the agendas of governments in relation to vaccines enter public discourse. This is particularly so as increasing numbers of vaccines are introduced, there are emerging opinions that children are being over-vaccinated, and not necessarily for their own good.

Negative ideas and conversations about vaccine risks spread rapidly through media and social media (Larson et al., 2013), but solid research that proves the effectiveness and safety of vaccines depends on scientific processes (Crawford et al., 2014) and cannot keep pace with

prevailing conjecture. This time lag between perception and evidence means there is much opportunity for the spread of misinformation regarding vaccines and vaccine safety. From the suburbs of any modern western city to bush clinics in remote settings, the use of media, social media, email, and cell phones results in rapid dissemination of negative news and "fake news" about vaccines, whereas countering this flood of negativity with good news about vaccines requires continuous medical and educational effort.

Maximizing vaccine safety leads to a smaller number of safety issues, and a more confident uptake of vaccines, and reduces the risk of the impact of future vaccine controversies. Vaccine safety is relies upon rigorous compliance to development and manufacturing standards, well conducted clinical trials,thorough assessment, licencing, control, and administration of vaccines. *Vaccine safety surveillance* is a vital component of safety effort, as it measures the effectiveness and safety of vaccines in the population at large (Chen et al., 2015).

2.2.1 Surveillance definition

Vaccine safety surveillance is inferred by observing rates of Adverse Events Following Immunization (AEFI). There is a standardized approach to describing AEFI, to enable scientific research and comparisons of AEFI data (Kohl et al., 2007). The World Health Organization (WHO) AEFI definition reads:

"Adverse Event Following Immunization (AEFI): This is defined as any untoward medical occurrence which follows immunization, and which does not necessarily have a causal relationship with the use of the vaccine. The adverse event may be any unfavourable or unintended sign, an abnormal laboratory finding, a symptom or a disease." (WHO, 2013).

As defined by the WHO, an AEFI could be any unfavourable reaction in a recipient, but also might be measured by abnormal laboratory findings, and longer-term results such as abnormal symptoms and diseases. As an example of long-term effects, almost 3 million Australian children were administered polio vaccines that were contaminated with a monkey virus in the 1960s, this has been linked to cancers in some of the recipients, particularly with mesothelioma (Cutrone et al., 2005).

2.2.2 Vaccine safety assessment

Vaccines are created from attenuated or dead viruses that induce the body to produce antibodies that protect against contracting the wild virus (Pulendran & Ahmed, 2011). Therefore, a range of mild and temporary reactions to a vaccine are expected: effects like swelling of an injection

site, a temperature, an upset stomach and the like. However, occasionally more severe AEFI can occur (Principi & Esposito, 2016).

Vaccine safety is assessed before and after licensure, in fact throughout vaccine development. Extensive testing and review before licensure focuses on the key areas of *safety*, *immunogenicity* (the effectiveness of antigen-induced immune response), and *efficacy* (the potential for a vaccine to protect from disease), culminating in three phases of clinical trials (Mcphillips & Marcuse, 2001).

2.2.3 Pre-licensure (clinical trials) surveillance

Once governing body approval is granted, three phases of clinical trials follow, which utilize increasing numbers of test subjects, with a goal of ramping up testing as the vaccine proves safe and efficacious at each phase. Despite the extensive testing these trials typically only discover expected and temporary side effects and reactions, including any interactions between the new vaccine and already utilized vaccines. Pre-clinical trials seldom detect unusual or rare adverse events, or delayed onset adverse events (Salmon & Halsey, 2016).

The use of controlled, double-blinded, and randomized trials ensures even and unbiased distribution of the vaccine throughout the test groups and safeguards the validity of any vaccine safety signals that may be detected. Even so, detection of rare and severe AEFIs during these clinical trials is unlikely, as these events arise only in a tiny percentage of much larger populations, which is only measurable during post-licensure surveillance (Lopalco et al., 2010). Furthermore, clinical studies are usually carried out on healthy subjects and within restricted age groups, and do not include individuals who are most at risk (such as those with medical conditions), and there is little additional evaluation of long-term or delayed onset health issues. Therefore, outcomes of clinical trials are not necessarily generalizable to the entire population (Salmon & Halsey, 2016).

Once clinical trials have been completed the results of these trials and additional steps require regulatory approval before a license is granted to distribute a vaccine to the general population (Marshall & Baylor, 2011).

2.2.4 Post-licensure surveillance

Phase IV trials are sometimes conducted by drug companies, where continued monitoring for safety and efficacy is needed. Internationally accepted guidelines provided by the WHO (WHO, 2010) are typically used by organizations conducting post-licensure vaccine testing, including the Australian TGA.

Vaccine safety surveillance continues in a variety of forms after regulatory approval. It is intended to identify serious rare AEFIs 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 (Chen et al., 2015). Surveillance also includes clinical practices in vaccine administration, to check that they adhere to the high standards required for safe delivery of vaccines. These systems can be categorised as passive, active, and a combination of both.

Passive surveillance systems

Passive systems typically rely on spontaneous reporting of adverse events by individuals, including vaccine manufacturers, physicians, health clinics, immunization programs, vaccine recipients and their caregivers (Varricchio et al., 2004). This kind of reporting is characterized as passive because it relies on voluntary reporting of AEFI (though manufacturers are obliged to report), rather than information being actively sought or extracted by health authorities. These systems are the main method of gathering Adverse Drug Reactions and have proven useful in early detection of vaccine and drug related safety (Clothier et al., 2017; Härmark & Van Grootheest, 2008). Examples of these systems in Australia are the Therapeutic Goods Administration (TGA) Adverse Drug Reaction System (ADRS) and the Adverse Events Following Immunisation – Clinical Assessment Network (AEFI-CAN) database for reporting of vaccine adverse events in Victoria and Western Australia. These systems have the advantages of being cost effective, having access to national data and enabling the formulation of hypothesis about new events (Isaacs et al., 2005). Among disadvantages of these systems are underreporting, and potential reporting bias (Pal et al., 2013). For routine immunization surveillance, passive systems remain the main method, and have proven to be particularly effective at detecting a large range of AEFI, most especially rare severe adverse events. Although these systems are the backbone of drug safety monitoring, there is a need to actively search for alternate data to get a more accurate picture of the quantity of possible adverse events.

Enhanced passive surveillance systems

Enhanced passive surveillance consists of a health authority actively seeking additional information, after receiving a passive surveillance report. Although passive systems are effective in identifying AEFI reports, data tends to be incomplete, so further investigation is warranted, particularly when a pattern of reports is identified. For example, Clothier et al. (2011) describe SAEFVIC as an online portal for accepting reports from both patients and

immunization providers after an AEFI. Hinrichsen et al. (2007) examined a system where the AEFI surveillance system references patient's electronic health records (EHRs) to determine likely AEFI, to then contact the patient's physicians via SMS and instruct them to make additional reports about the patient. Lazarus et al. (2009) discuss a similar approach whereby clinicians are prompted to consider whether vaccination might be the cause of patient's condition and where their further reporting to a passive surveillance system is facilitated.

Active surveillance systems

Active surveillance systems make an active effort to search, identify and collect AEFI (Griffin et al., 2009). There are major disadvantages of passive systems: underreporting and incomplete data, a preponderance of reporting on already known adverse events, and a tendency to report on coincidental events. Consequently, active AEFI surveillance techniques have emerged as complementary approaches to secure more comprehensive and clinically reliable data (Harpaz et al., 2016).

Prospective surveillance is a type of active surveillance where a targeted group of AEFIs are monitored. An example is the Canadian Immunization Monitoring Program, ACTive (IMPACT) surveillance system, which uses active paediatric hospital-based surveillance (Ja et al., 2014). A monitoring nurse and supporting personnel in targeted Canadian paediatric hospitals collect data on children admitted with symptoms that are typically encountered post-vaccination. Australia has also set up an active surveillance system based on the IMPACT model, which is called Paediatric Active Enhanced Disease Surveillance (PAEDS) and monitors medical encounters at major paediatric hospitals across Australia for potential cases of AEFI (Zurynski et al., 2013).

Another form of active surveillance is through data record linkage. This approach endeavours to link vaccine history data with data coming from a variety of other health records such as electronic medical records and health insurance claims, where known AEFI and specific Health Outcomes of Interest (HOI) are targeted for the data linking process. A leading example is the North American Vaccine Safety Datalink (VSD) project (McNeil et al., 2014), maintained by the Centre for Disease Control which conducts active surveillance of vaccine safety using large linked databases. VSD enables the study of causality between the adverse event and the vaccination. The main method for confirming possible vaccine reactions is to link a medical encounter's likely vaccine event-related diagnostic codes to an actual vaccination. VSD can be used in two ways: Either by looking at historical data to establish if some adverse events are more common after receiving a specific vaccine; or to monitor

incoming data to establish if certain events are more prevalent among vaccinated people. VSD uses large-scale distributed data networks which allows almost near real-time vaccine safety surveillance, with most data being updated on a weekly basis. Another example is the US Federal Drug Agencies Post-Licensure Rapid Immunization Safety Monitoring (PRISM) system (Baker et al., 2013). This system uses large claims-based distributed databases, which allows monitoring of vaccine safety on a large population of more than 25 million.

Mesfin et al. (2019) performed a systematic review on the use of EHRs for post-licensure vaccine surveillance and found that they are increasingly used for near real-time AEFI detection. They observed that there are opportunities for checking the utility of non-coded patient encounters for collecting additional AEFI surveillance data, and suggest the use of telephone helpline phone conversations, which they explore in a later study (Mesfin et al., 2020).

User-generated data

Established passive reporting systems increasingly provide online public reporting interfaces. Although non-professional reporting on health issues are of a different quality from professionals reporting, it is uniquely valuable — as it is the patient's perspective and voice and can in fact be quite reliable and objective at the same time, provided it can be interpreted (Seifert et al., 2017). For instance, a study by Krska et al. (2011) shows that patients usually feel confident to identify and describe their adverse drug reactions when reporting them to the Yellow Card Scheme web site run by the UK Medicines and Health Care Products Regulatory Agency. It has been found that their reporting aligns with that of health professionals, and the authors conclude that patients' reports are reliable and should be taken seriously (Krska et al., 2011).

A study by Clothier et al. (2014) showed that even though reporting by patients on the AEFI-CAN (Adverse Events Following Immunisation – Clinical Assessment Network) site is considerably less than professionals' reporting, patients experiencing *serious* AEFIs are more likely to report them than health professionals, which indicates the value of patients' initiatives in data collection. This result is also confirmed through a study done by Karapetiantz et al. (2018) reporting that patient-reported reactions in forums, although less informative, contain more unexpected reactions.

Social media data has become an additional and widely used online source of data for public health research (Conway et al., 2019). There are many examples of social media data being analysed in communicable and non-communicable disease monitoring, for drug use and abuse

studies, for measuring mental health issues, and for the impacts of health policies including vaccinations (Milinovich et al., 2014).

To conclude, this subsection introduced the importance of vaccine safety and how vaccine surveillance for adverse events following immunisation (AEFI) is a vital component of ensuring vaccine safety. Traditional vaccine surveillance reporting systems were described and problems with these systems were explained — including underreporting and deficiencies in timeliness and coverage — which has led to efforts to expand possible reporting sources. Social media users' self-reporting on health-related issues was explained as an emerging complementary data source for timely public health surveillance, which is explored in detail in the next section.

2.3 Social media data sources for public health studies

For public health studies, there are two major categories of social media platforms. General-purpose domains such as Twitter, Facebook and Reddit; and domain-specific platforms such as PatientsLikeMe, DailyStrength and drugs-forum. The domain-specific social media forums provide data that focuses on specific health issues and interests. In contrast, general-purpose social media conversations cover a broad range of topics and are more suitable for discovering current trends and common subjects. Topics like the COVID-19 pandemic and vaccines in general are discussed by everyone and are well mentioned in general-purpose social media.

Twitter, Facebook and Reddit are the most used social media sources for research in public health (Dol et al., 2019; Singh et al., 2020; Tang et al., 2018). The privacy and data sharing policy of these platforms impacts their popularity in research. Facebook, although the largest social network is very seldom used in research, mainly due to its strict privacy policies. Reddit has been a source of data for public health studies — however, it is used a lot less than Twitter. Compared to Reddit, the frequency of information dissemination in Twitter is much higher, the interval between posts is measured in minutes rather than hours. Additionally, the short text requirements of tweets encourages focused conversations that use more prevalent terms (D. Choi et al., 2016). According to Priya et al. (2019), tweets reduce the articulation of biased and extreme views, compared to Reddit posts. They also find that Twitter conversations can gather a greater momentum than Reddit posts, which can result in a news event being sustained for much longer in Twitter, and that the spread, availability, and low inter-arrival times in Twitter posts make them very suitable for real-time monitoring, including emergencies.

Amongst the general-purpose social media platforms Twitter is most widely used as an additional source of data in the pharmacovigilance domain and in public health research (Edo-

Osagie et al., 2020; Gupta & Katarya, 2020; Singh et al., 2020). Its real-time nature, high volume of messages and public availability makes it a suitable alternate source of data for surveillance of public health (Lardon et al., 2018).

2.4 Surveillance using social media

Surveillance as a major component of public health is defined as the continuous systematic collection, analysis and interpretation of health data for planning public health actions (B. C. K. Choi, 2012). Social media platforms have become a valuable data source for public health research and monitoring. Existing research in public health-related social media use can be categorized into two major groups. One group is concerned with the use of social media as a platform for sharing knowledge and information communication; the other has to do with using social media for knowledge discovery and building predictive models (Zhou et al., 2018). In a systematic review performed by Sinnenberg et al. (2017) a taxonomy of Twitter use in health research is clustered into six categories. Four of the categories, including content analysis, surveillance, engagement, and network analysis, belong in the knowledge discovery group; the two other categories, including recruitment and intervention, belong in the information communication group. This research falls into the second group of knowledge discovery, specifically the discovery of vaccine surveillance information from social media.

2.4.1 Disease surveillance

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 (Paul & Dredze, 2017). Social media is used as a data source for surveillance of various disease, including:

Infectious disease

Influenza and influenza-like illnesses (ILI) comprise a large proportion of infectious disease surveillance on social media. Several studies have used social media data and established that the accuracy and effectiveness of detection of ILI could be improved by using social media data (Lampos et al., 2017; Signorini et al., 2011; Velardi et al., 2014).

Other surveillance research demonstrating the usefulness of social media includes the monitoring of Ebola in Nigeria using Twitter data (Odlum & Yoon, 2015); Zika surveillance combining traditional disease surveillance techniques along with search logs, social media, and

news report data (McGough et al., 2017); Cholera outbreak tracking using Twitter and news media data (Chunara et al., 2012); and H1N1 (swine flu) surveillance using Twitter data (Signorini et al., 2011).

Non-Infectious disease

Examples of studies which have explored user-generated content for non-acute chronic disease include: Early detection of pancreatic cancer from user search log data (Paparrizos et al., 2016); establishing the effect of change of weather on fibromyalgia patients (Delir Haghighi et al., 2017); and evaluating Twitter conversations for discussions that might indicate links between psychological characteristics and risks of heart-disease mortality (Eichstaedt et al., 2015). Twitter has also been investigated as a resource for allergy surveillance and was proved useful for insight generation and early detection of hay fever (Rong et al., 2019) and thunderstorm asthma outbreak (Joshi et al., 2020).

Other illnesses and health conditions

Other illnesses such as foodborne disease (Sadilek et al., 2017), mental health (Coppersmith et al., 2014) and obesity have also been studied using social media (Ghosh & Guha, 2013).

2.4.2 Adverse Drug Reaction detection

Pharmacovigilance is defined as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" (World Health Organization, 2002). In a comprehensive survey done by Karimi et al. (2015) on text and data mining techniques for Adverse Drug Reactions (ADR), social media and online sources such as search logs and forums are identified as one of the main data sources used for discovery of ADRs.

Many researchers have successfully established the usefulness of social media as a pharmacovigilance source. Research in ADR detection is mainly focused on either binary classification of sentences with mention of an ADR, and/or extraction of ADRs. Lardon et al. (2015) 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. Sarker et al. (2015) reviewed articles published from 2010 to 2014 on automatic pharmacovigilance utilising social media. They noted a shift of emphasis in this area of research, from exploratory studies to more structured approaches. This has resulted in

an increased interest in using supervised machine learning techniques, which require 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 (Klein et al., 2020; Sarker & Gonzalez-Hernandez, 2017; Weissenbacher et al., 2018; Weissenbacher & Gonzalez-Hernandez, 2019). The organizers provided participants in the ADR classification task around 25,000 of annotated tweets for training, and 5000 for validation — and these datasets remain available upon a request to the organizers.

2.4.3 Vaccine Adverse Event detection

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 (2017), 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 (Sarker & Gonzalez, 2015) , for example an annotated set of 10,822 tweets which the author assessed. The text of these tweets needed to be downloaded, and the author was able only able to obtain 6,670 of them — there were no vaccine adverse events. Downloading the accessible tweets of the larger unlabelled dataset only produced 158,028 tweets. Around 100 posts that mentioned influenza(flu) shots were all within Tamiflu discussions, and only 80 posts contained a vaccine mention. There were no AEFI mentions.

A recent study by J. Wang et al. (2019) specifically addressed the challenge of flu shot adverse event detection. Their concern was to find social media posts that contained specific mentions of adverse events that were being experienced by users that were known to have recently had a vaccination. In their work they emphasized 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 mentions that were definite adverse events, as specified by a domain expert. Although this approach somewhat focussed the data annotation process, it still required a few 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 (J. Wang et al., 2018). Consequently, their work has interesting techniques for integrating the language of formal reports with that of social media data. The focus on identifying specific and verifiably adverse events means that their goals are like those of studies on adverse drug reaction detection, where the emphasis is on distinguishing significant drug reactions.

This research is similar in scope to ADR and AEFI detection, but rather than focussing on detecting specific known events, the research goal is to detect the kind of language that is used in relation to the experience of a potential adverse event, so that the data can give insights into the general trends of such conversations, but is also capable of capturing any kind of adverse event related text – and to not be constrained to detecting specific events. The next section explores an area of research more related to this work, which is using social media for personal health mentions detection.

2.4.4 Personal health mention detection

One area of research in social media language processing is personal health mention detection. This deals with identifying posts which have a mention of a health condition and the person affected by it. Joshi at al. (2019) have published a survey of text-based health mention detection datasets, approaches and evaluation methods. Yin et al. (2015) collected tweets across 34 health conditions and showed that combining posts from a number of health topics into four categories can train a personal health mention classifier that performs better (with 77% accuracy) than a classifier trained on a single health issue. They used a Multinomial Naïve Bayes classifier but most recent work in this domain has used deep learning algorithms, either in feature engineering or as a classification technique. Karisani & Agichtein (2018) developed approaches for configuring word embedding representations which assisted the discovery of the most effective features to use in a statistical classifier and reported an improvement over previous methods using similar data. They used two labelled datasets of 3,000 and 7,000 tweets. Iyer et al. (2019) used the same datasets and developed a deep learning approach with augmented features for idiom detection which showed further improvement in personal health mentions detection. Wang et al. (2021) were able to access 5,288 tweets from this dataset and used CNN-based classifiers with word embedding features to observe the impact of the number of training samples on the performance of models. In another study, Jiang et al. (2019) collected around 22 million tweets on 103 medicines, annotating around 12 thousand tweets, and confirmed that word embeddings as inputs to a long short term memory neural network

(LSTM) consistently performed better than traditional classifiers with bag of words plus engineered features.

Personal health mention detection has also been defined as a task in SMM4H (Weissenbacher & Gonzalez-Hernandez, 2019). Five teams participated in the task and the systems have been evaluated on their generalizability across different health domains. The architectures of all the participated systems were based on Transformer language models.

This area of research, although explored for applicability to various diseases, has however not been studied in relation to vaccines. The next section discusses areas where most of the vaccine-related social media research has been focused.

2.4.5 Monitoring of vaccines and vaccinations

Most of the research in social media monitoring of vaccine and vaccination mentions is in understanding and analysing sentiments, opinions, behaviour, and attitudes. Salathé & Khandelwal (2011) analysed vaccine sentiments in Twitter posts about the influenza A (H1N1) vaccine. 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. Du et al. (2017) assessed Twitter sentiments towards human papillomavirus (HPV) vaccines to understand public opinion and concerns.

Other researchers have done more in-depth analysis to find the themes and sentiments of social media posts about a vaccine during an outbreak. Radzikowski et al. (2016) studied Twitter posts to understand public attitude toward the measles vaccination during the US measles outbreak in 2015; Mollema et al. (2015) analysed tweets to address public concerns during the 2013 measles outbreak in the Netherlands.

Automatic classification of tweets mentioning vaccine behaviour was one of the tasks in the third SMM4H shared task (Weissenbacher et al., 2018). The task was a binary classification of tweets that indicate the intention to receive flu vaccine. Ten teams participated in this task and the winning team used an ensemble of statistical classifiers with task specific features combined with rule-based and deep learning models (Joshi et al., 2018).

Table 1 provides a summary of some of these studies, which illustrates that they are mostly concerned with vaccine sentiment or opinion rather than with AEFI detection.

Table 1: Summary of vaccine-related studies

Study	Source	Data size	Aim	Method
(Salathé & Khandelwal, 2011)	Twitter	477,768	Measure the spatio-temporal sentiment towards influenza A(H1N1) vaccination / 3 categories	Naive Bayes and the Maximum Entropy classifiers
(Bello-Orgaz et al., 2017)	Twitter and WHO data	761.924	Measuring the potential influence of vaccine opinions based on the variation in the coverage rates	community detection algorithms
(Huang et al., 2017)	Twitter and CDC data	1,007,582	Track vaccine attitudes and behaviours on Twitter and infer vaccine-related intentions from Twitter messages, focusing specifically on the influenza (flu) vaccine / Binary	SVM, Multinomial Naive Bayes, RandomForest
(Radzikowski et al., 2016)	Twitter	669,136	Analyse themes and relations that make up the discussion about vaccination in Twitter	community detection, network visualisation & analysis
(Du et al., 2017)	Twitter	184,214	To understand public opinion about HPV vaccines	SVM models
(Huang et al., 2018)	Twitter	1,124,839	To measure influenza vaccination uptake through Twitter and track vaccine attitudes and behaviours	convolutional neural network
(Lama et al., 2019)	Reddit	22,729	To determine the topics of discussions on HPV vaccine related messages on Reddit	Latent Semantic Analysis
(Wang et al., 2019)	Twitter	3,139 users and their 90,185 tweets	adverse event-indicative messages from known vaccine recipients	LibShortText to identify users; Semi-supervised Multi-instance (SSM) models to identify AEFI
(Deiner et al., 2017)	Facebook and Twitter	58,078 Facebook posts and 82,993 tweets	To measure public opinion towards measles vaccination	social media analytics platform

This subsection reviewed the existing studies which utilize social media as a data source for public health surveillance. The analysis revealed that the use of such data is established, and more advanced mining techniques are being developed. However, the review noted that there is a relative deficit in vaccine adverse event mention research, with investigations of vaccine and vaccination-related social media posts characterized as mostly concerned with sentiments,

attitudes, and opinions. The next section describes the methods used in social media monitoring, which includes the applicable techniques for this study.

2.5 Social media data processing

Social media data are textual and range from news and stories expressed in a structured and more formal writing style through to anecdotal and idiosyncratic personal accounts and expressions. Social media platforms vary in their purpose and this is reflected in type of writing that can be found in them. For instance, Facebook allows people to write at length and to conduct ongoing personal, and even private, conversations; Twitter's limited individual message lengths and its public nature encourages declarative language with the intent of rapid information dissemination; blogs, creative writing and news forums can constrain writers to specific structures, subjects and language and can even have quality controls over submissions.

Many social media platforms publish aids for accessing data from them in the form of Application Programming Interfaces (APIs), most however limit the amount of information that can be freely accessed. Accessing and processing the texts of social media platforms is non-trivial and there are many resources and much research dedicated to these tasks, which this investigation makes use of, and is explored in this section.

2.5.1 Social media data collection

The first task when using mining social media is to collect data. Although some existing data may have been collected and curated, typically a researcher must collect their own data to meet their specific requirements. Usually this involves using the social media API to search or filter the data that is continually streamed online from the social media source. There are a variety of methods employed by researchers to filter the collection of social media posts of interest. Some of those include keyword and phrase selection, keyword selection with phonetic spelling and direct selection of users. The simplest method for collecting relevant content is to filter for social media messages using search queries containing certain keywords or phrases relevant to the task. Examples include research on emotion classification of tweets before and after the Ebola outbreak of 2015 (Ofoghi et al., 2016), and research on community opinion clustering about HPV vaccines by searching for "hpv" and "vaccine" combined, and variations of "Gardasil" (Surian et al., 2016). Pimpalkhute et al. (2014) describe a more generalizable approach, by generating phonetic spelling versions of the search term and its most frequent variations and conducting searches phonetically. For example, phonetic spelling of drug names has been used to target relevant social media posts. Another way to filter pertinent data is to

focus on posts that come from specific forums or blogs, or that include mentions of hashtags and other specific references that help to narrow the range of posts, or that come from people who are followers of particular subjects or brands. An example of a study using direct selection of users is the work done on automatic detection of e-cigarette use by filtering out the content based on the e-cigarette brand followers and hashtags (Aphinyanaphongs et al., 2016). Another approach which removes the bias introduced by keyword searches is to not have an initial keyword filter and collect everything for a given period or location (Cameron et al., 2012).

2.5.2 Text pre-processing

Texts collected from social media sites is normally processed by machine learning (ML) models — these automate the task of exploring and developing an understanding of the textual information, which would be virtually impossible to do manually. Typically, texts require some form of tokenization and the addition of features, then they must be converted into a numerical form for consumption by ML models. The following sections describe some of the NLP techniques used and their preparatory requirements.

Bag-of-words

To a human reader text has a logical sequence and the clear meaning of text is gained through reading it in sequence, and often with an understanding of meaning gained earlier in the reading of a text and even with consideration for the purpose of a text and an anticipation of its conclusion. Reading and comprehension are highly complex and sophisticated processes, and even the most advanced computer technology cannot truly emulate them. However, for determining the overall subject matter or emotion of a text there are computer-based approaches to understanding text that work reasonably well. Statistical techniques applied to the word counts are used to make decisions about document meanings. This is commonly referred to as the "bag-of-words" approach because words are just counted as being members of an unordered collection, as if they were loosely bagged up rather than laid out in a significant order. The bag-of-words approach is typically the first type of computer based natural language processing that is performed — it is surprisingly enough for many tasks and is used as a benchmark for more sophisticated techniques. A bag-of-words representation is the core representation utilized in all the search engines (Zhai & Massung, 2016).

N-grams — preserving some word relationships

The bag-of-words approach does not attempt to understand phrases, all words are just collected together and counted and analysed, however it is very often the case that words are repeatedly found together as phrases, and that the phrases have their own significance which probably should be preserved. Therefore, when constructing the vocabulary of a corpus of documents, the practitioner should consider preserving phrases, to compensate for the loss of meaning enforced by the bag-of-words approach.

Word n-grams are sequences of words which are comprised of the most likely words that happen in each text segment (Jurafsky & Martin, 2007) and are one of the most informative and commonly used features in text processing tasks (Sarker et al., 2017). N-grams are a basic representation of word order and relationships in that they create new words from commonly encountered sequences of words — for example if the phrase "black sheep" was often encountered in a document collection then a bigram (2-gram) would preserve this as the new word "black_sheep", which would have a quite different significance to the individual words "black" and "sheep".

Lemmatization and Stemming — reducing word differences

It can help a bag-of-words approach if differences between words are minimized by reducing them to their root forms. Lemmatization uses a dictionary of linguistically correct bases of words (the lemma) and can also consider the word context when deciding an appropriate lemma, and so preserves a natural and readable form of the common word. The lemma of a word might be a representative — for instance "good" is considered by some implementations as the lemma of "better". Stemming however just chops off the ends of words to a representative stub and without any consideration of context, and so can be quite unreadable — however it is much faster to stem a document, as a dictionary lookup and accounting for word context is not required. Some root words might be the same using either approach. for example, the words "walks", "walked", "walking" are all reduced to "walk" by both approaches.

Representing words and documents as vectors

Machine learning algorithms need text to be converted to numbers to process them. A standard ML approach is to first assign each word a numeric position in a dictionary that contains all the assembled documents' words. Then, for each document, to represent its words by referring to the dictionary positions of the words that are present in the document. This can be

represented as a vocabulary-length vector, with counts of only the currently present words being non-zero, at their position in the vector. As the length of these vectors is as large as the dictionary vocabulary and the vector consists of zeros unless a word is encountered in the document, they are known as sparse matrices. Normally however, the sparse matrix is represented by pairs of index positions and word counts per word in the document, which avoids having to represent the words that are not currently present.

TF-IDF: Words in document context, and in corpus context

The simplest technique to implement bag-of-words is to treat text as a collection of words and to just collect the words and count their frequency. This technique is called Term Frequency (TF). In its simplest form, where t denotes a term and d denotes a document, the formula of the TF calculation is:

$$TF(t,d) = \sum_{x \in d} fr(x,t)$$

$$fr(x,t) = \begin{cases} 1, & if\ x = t \\ 0, & otherwise \end{cases}$$

Inverse Document frequency (IDF) can be used with TF to give more importance to specific terms by downgrading frequently used terms. Where D is the corpus and $|\{d: t \in d\}|$ is the number of documents containing term t, the IDF formula is defined as:

$$IDF(t) = \log \frac{|D|}{1 + |\{d: t \in d\}|}$$

That is, the more a term appears in a document the less discriminative it is (Jones, 1972).
This combined approach to text feature extraction is called Term Frequency-Inverse Document Frequency (TF-IDF). The TF-IDF formula is defined as follows:

$$TF - IDF(t) = TF(t,d) \times IDF(t)$$

Although TF-IDF can model some of the relative semantic importance of the words in a document, it cannot however encode any information about term similarities (Kowsari et al., 2019).

A major shortcoming of these methods is that they do not intrinsically contain any information about words such as similar or opposite meanings or contexts. Any word significance and relationship are solely derived by the model as it processes the words in the

context of the individual documents and the document's similarities and differences with other documents in the dataset. Alternative *word embeddings*, which embody intrinsic relationships between words, have been subsequently developed — these are used by neural networks and to some extent can be utilized by classic machine learning models, this is explored in the Dense Word Vectors section below.

Dense Word Vectors

When using one-hot encoding, words do not embody any semantic value, they are learned about by the classifier in terms of their frequency of appearance in the document collection. If a new word is introduced the model cannot use it, as it has not previously learned about it. However, if a word could be represented with a numerically expressed value that placed it in a context of similar words then it could be understood by a suitably engineered model as belonging in that context, so the model can gain additional "contextualized" understanding of words, as well as semantic ones (Chang & Chen, 2019).

Providing computationally efficient representations of words which can capture word similarities has been the subject of much research (Mikolov et al., 2010). The resulting approach is to use "dense" vectors, named word embeddings, which represent (embed) words in a continuous vector space, and which are obtained from a shallow two-layer neural network that can discover semantic relationships. Words are evaluated in their context over many thousands of examples, and a three-dimensional vector space is constructed to place words into relationships with one another. Each word is represented by a vector of large floats, typically between 100 and 300 numbers per vector, where the value of individual numbers at each position in the vector are related to equivalent values used by other words — with the result that words that inhabit similar contexts in the corpus are positioned near to each other in the vector space. At the very least these vectors can provide a more succinct word representations than a sparse matrix can (e.g., 100 numbers per word instead of possibly thousands of zeros and a single number one), but the best use of them is to provide an excellent starting point for neural networks for understanding and clustering words of similar value, and they can even be used by classic machine learning models to infer document similarity via the combined values of words in a document. The sum of the numbers in the individual vectors place them into positive or negative space, and no individual word's values sum to a very large number, therefore the value of any word does not carry any particular significance apart from its place in the vector space — which has the same effect of eliminating unwanted bias as the one-hot encoding approach.

Neural network approaches for modelling these distributed word representations include predictive CBOW (Continuous Bag-of-Words) and Skip-gram models (Mikolov, Chen, et al., 2013), and co-occurrence matrices (Pennington et al., 2014). A CBOW model predicts the current word based on the context of surrounding words, maximizing the probability of a target word by examining its context. The skip-gram model predicts words within a certain range before and after the current word, suggesting the most likely surrounding words. A co-occurrence model counts how frequently a word appears in a context. There are various implementation of these, the most well-known are Word2Vec which can use either CBOW or Skip-gram approaches (Mikolov, Sutskever, et al., 2013), and GloVe (Global Vectors for Word Representation) introduced by Pennington et al. (2014), which uses the co-occurrence approach.

These various text pre-processing approaches all have the goal of rendering the texts in a numeric form suitable for consumption by machine learning text classifiers, which are described in the following section.

2.6 Machine learning methods in text processing

For the downstream processing of social media text in a health-related context most studies can be categorized into two main groups, one consisting of text classification and content analysis tasks and the other of information extraction (IE) and normalization tasks (Gonzalez-Hernandez et al., 2017). In terms of the task of surveillance, these can be thought of either attempting to find specific indicators of the subjects of interest — using content analysis and/or text classification; or extracting medical terms or lay equivalents (IE) and mapping them to medical ontology identifiers (normalization), to discover what is being said. In the next section we review two categories of studies in each of these domains that related to the main research question of this study namely medication safety as a problem of information extraction and personal health mention detection as a classification area.

Examples of well-established social media text classification and content analysis tasks in the health domain include sentiment analysis, emotion analysis (Ofoghi et al., 2016), real-world trend detection (Sun et al., 2017). Conversely, Adverse Event detection (Freifeld et al., 2014; Huynh et al., 2016) is an Information Extraction task.

While research into information extraction and medical terms capturing from medical texts is an established area (Khademi et al., 2015), performing IE and normalization on social media texts is relatively less developed, complicated by the lexical complexities of both the social media and medical domains (Baldwin et al., 2015).

One way of categorizing the approaches used in these tasks is based on their practical requirements and purpose — as being supervised, unsupervised, or semi-supervised. These categories describe the approach to training the models but are also indicative of the goal in using the approaches. Supervised learning is when the outcome is known beforehand, so the model is guided to accomplish the goal, for instance to classify a text. Unsupervised learning is when the model is to discover what is significant in data without any supervision, for instance to discover latent topics in texts. Semi-supervised learning is when the model should be guided to accomplish a task but lacks the data needed for fully supervised process, instead the model is assisted to discover the relevant data.

Supervised methods

Supervised machine learning methods are typically used to match words or overall subject matter of a text to specific categories. That is, with supervised tasks we know in advance how the text needs to be categorized, and examples of the texts are labelled with these categories and used to enable the ML models to *learn* how to recognize these sorts of texts. An example is Lamb et al. (2013) work, where after acquiring the initial Twitter dataset with a targeted search a ML model is trained to filter the data to health-related tweets and then use an additional model on that dataset to find flu-related tweets. Supervised ML is often referred to as *classification*, as the models are trained to separate texts according to labels or *classes*. A model needs to be able to perform its classification task in a sufficiently generalizable way so that subsequent but different texts given to the model can be correctly categorized based on their similarities to the original texts. As supervised ML relies on having labelled texts to learn from, it is most often a downstream process after texts have been understood and grouped sufficiently to be amenable for the labelling required for supervised ML models. This preliminary grouping is typically a task suitable for unsupervised machine learning.

Semi-supervised methods

Semi-supervised methods require a small amount of labelled data to help guide the model towards discovering patterns in the data (Zhu, 2005). A semi-supervised approach bridges between unsupervised and supervised learning. Lee et al. (2017) used a variety of unlabelled random tweets to pre-train several semi-supervised Convolutional Neural Networks and reported improved performance in supervised classification methods over models trained only on labelled data for the task of adverse drug reaction detection.

Unsupervised methods

Unsupervised methods do not require labelled data and use clustering to group data into categories, but the categories are not known beforehand. Basic unsupervised techniques include K-means clustering (Han et al., 2011), where similarities in vectorized texts are identified to enable clusters of words and thereby documents containing them. The next section examines Topic Modelling in depth, which is one of the most widely used methods for analysing large amounts of social media data and utilize more sophisticated unsupervised text clustering methods.

2.6.1 Topic modelling

Topic models use *dimension reduction* to uncover the common themes that convey similar semantic meaning in a corpus of documents (Blei et al., 2010). Topic models may take a non-probabilistic approach such as used in *singular value decomposition*, or use probabilistic techniques which assume that each document can be represented by a distribution over topics and each topic by a distribution over words (Barde & Bainwad, 2017). Simply put, a topic encapsulates its most likely words, compared to other topics; and a document can be understood by how its words are most likely to belong to specific topics — consequently, topic models can be used in classifying and summarizing documents. There are several methods of probabilistic topic modelling in the literature, one of the most popular methods is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is a Bayesian topic model. LDA models have been widely used in health-related social media mining (Benton et al., 2016; Paul & Dredze, 2014).

Latent Dirichlet Allocation (LDA)

LDA based models are generative probabilistic models and are presently considered a state-of-the-art method for topic derivation (Nugroho et al., 2020). They work on the assumption that each document can be represented by distribution over topics and each topic by distribution over words. LDA models corpus D, containing M documents, based on following the generative process of (Blei et al., 2003):

1. Randomly choose $\mathcal{T}$ topic distribution, $\beta_t \sim Dirichlet(\lambda_\beta)$
2. For each document $d = (w_{d1}, w_{d2}, \ldots w_{dn})$:
    a. Randomly choose a distribution over topics, $\theta_d \sim Dirichlet(\lambda_\alpha)$
    b. For each of the N words in document d, w_{dn}:
        i. Randomly choose a topic $z_{dn} \sim Multinomial\ (\theta_d)$
        ii. Randomly choose a word $w_{dn} \sim Mutinomial(\ \beta_{zn})$

Then the probability of a corpus is calculated as follows:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) P(w_{dn}|\, z_{dn}, \beta) \right) d\theta_d$$

The parameters α and β are to be sampled once in the process of generating a corpus and therefore are corpus level. The variables θ_d are document-level, and the variables z_{dn} and w_{dn} are word-level variables.

Topic modelling on short text

Topic modelling on short text has particular challenges. This is due to content sparsity and limited context, and in social media the further challenges of informality of language and noise. Three major approaches have been used to deal with short texts: One is to aggregate short texts into pseudo documents based on a common contextual feature such as author, location, or use of hashtags, and then to apply standard topic models to the aggregated document (Bicalho et al., 2017; Quan et al., 2015). Another approach is to use global word co-occurrence based models, where topics are learned from a collection of all the words in a corpus (Cheng et al., 2014; Zuo et al., 2016).

The third approach is to make an assumption that each document can be described by one topic. Either the dominant topic of a potentially multiple-topic document can be assumed as the one topic, or a topic modelling algorithm that deduces only one topic per document can be used. The latter approach has been handled by topic models known variously as mixture of unigrams and Dirichlet Multinomial Mixture (DMM) models (Nigam et al., 2000). The assumption of one topic is a very reasonable one for short texts and means that topic identification can be consolidated and compensate somewhat for the content sparsity problem. Zhao et al. (2011) applied the one topic per tweet model to a Twitter corpus of 1.2 million tweets and compared it with LDA and Author topic models using human judgment on topic coherence, and found that it ourperformed other models, giving more meaningful top topic words. Yin & Wang (2014) compared DMM with some other clustering techniques on a corpus of tweets and found that the inferred topic are more complete and heterogenous, and the models were faster to converge than the other techniques. Mazarura & Waal (2016) used measures of topic coherence and topic stability to show that DMM outperformed LDA on a tweet corpus. Over the years variations of DMM models (which incorporate word embeddings into DMM) have been proposed and proved promising results on short text topic modelling, some examples

include, LF-DMM (Nguyen et al., 2015), GPU-DMM (Li et al., 2016), GPU-PDMM (Li et al., 2017), ULW-DMM (Yu & Qiu, 2019).

In a review by Qiang et al. (2019) eight topic modelling approaches from the three categories of Self- aggregation, DMM, and global word co-occurrences based methods are compared on six different datasets including Twitter data. On Twitter dataset, DMM based models outperformed other models in classification accurency and topic coherence and purity.

Topic modelling in social media text mining

One main use of topic models is for exploratory analysis. Prier et al.(2011) used LDA to discover valuable tobacco-related themes from Twitter conversations. Ghosh & Guha (2013) combined LDA and GIS information to extract common and important obesity-related topics from Twitter in the USA. Paul & Dredze (2011, 2014) extended LDA to develop the Ailment Topic Aspect Model (ATAM), which demonstrated the possibility of automatically identifying health topics from a corpus of Twitter messages - topics included influenza and allergies.

Topic modelling is also used in classification tasks, in two major ways. One is to learn features based on unsupervised topic models and then to use the learned features in classifiers, either as main or additional features. As topic models identify the associations of words and so encapsulate semantic features of the data, including topic model-based features can improve classification results. The addition of topic modelling-based features has increased the performance of a binary classification task for classifying Twitter posts containing adverse drug reactions (Jonnagaddala et al., 2016), in detection of cyberbullying in social media text (Van Hee et al., 2018), and tweets sentiment analysis (Palogiannidi et al., 2016).

The other classification use of topic models is to directly act as a classifier, these methods are called supervised topic models. Supervised Latent Dirichlet Allocation (SLDA), introduced by Mcauliffe and Blei (2008), is the most popular supervised topic model to date (Dai Nguyen, 2019). These models are particularly favourable over other classification methods due to their ability to discover the underlying structure of words associations. SLDA is used for mining Twitter in identification and analysis of traffic incidents (Gu et al., 2016) and proved that social media can be a potential complementary source for incident reporting. Supervised topic models have also been used to detect linguistic signals in depressed individuals language (Resnik et al., 2015).

Another use of topic modelling is to speed document annotation (Poursabzi-Sangdeh & Boyd-Graber, 2015) and assist with automatically augmenting training data. Yang et al.(2015)

used topic models for a dimension- reduction mechanism and to augment training data used for building a robust classifier to identify adverse drug reactions in social media.

Topic modelling evaluation measures

Topic modelling evaluation measures can be categorized either as intrinsic or extrinsic. Intrinsic measures do not rely on annotated data and are mostly to evaluate the generalizability, quality (interpretability) or predictive power of topic models. These techniques can be manual and have human input in the loop, for example word/topic intrusion introduced by Chang et al. (2009); or can be automatic, such as perplexity (Wallach et al., 2009) and topic coherence (Newman et al., 2010). Qualitative intrinsic measures emulate and automate human input and the topics identified by them must be consistent with human understanding. When developing the coherence measure Newman et al. (2010) designed it to be coherent to humans, and asked humans to judge whether the learned topics of their models were both interpretable and associated with a single sematic concept. While the best arbiter of model suitability is the human practitioner, automated measures offer scalability (Dai Nguyen, 2019).

Extrinsic measures do require annotated data and generally evaluate topic models suitability for performing tasks such as classification. Extrinsic measures are suited for making comparisons between topic modelling methods; some major examples include purity (Y. Zhao & George, 2001), pairwise f-measure, and normalized mutual information (Manning et al., 2008).

The most useful evaluation measures are those that are applicable to the task that topic modelling is being used for. For instance, if the goal is to classify data then standard F-Scores will be applicable, and intrinsic measures should only be used for initial model evaluation. The most appropriate metrics depend on the application domain, and therefore there is no single best measure (Dai Nguyen, 2019).

Document clustering techniques are vital for making sense of the vast amount of wide-ranging data encountered in social media, and their outputs of homogeneous groups of documents are then amenable for the further machine learning task of classification, which is nowadays dominated by the neural network-based approach known as deep learning.

2.6.2 **Deep learning**

The underlying machine learning architectures for NLP tasks range from statistical modelling based on the volume of data (Gonzalez-Hernandez et al., 2017), to probabilistic modelling based on the frequencies of certain words, including combinations of words (Yala et al., 2017),

then to the use of word embeddings and deep learning (Miotto et al., 2017), through to fully fledged language modelling (Peters et al., 2017).

When using machine learning (ML) models, a practitioner must assemble increasingly refined representations of text as inputs for the models, by making iterative manual adjustments to the data until the optimal data structure for the model's capacity is determined. At the same time, the model's settings or hyperparameters must also be adjusted to best handle the inputs. These *input features* need to be engineered by system experts; the model just performs numerical optimization on the prepared data to derive the clearest conclusion about what the data represents. It can be said that much of the learning in this situation is performed by the human expert, who is refining the input features and model hyperparameters after feedback from the ML model's processing of them.

Deep Learning (DL) however, hands over learning about the data to the ML model — data needs minimal preparation but instead requires sufficient examples for the computer model to create its own internal representations from the data and to simultaneously tune itself. Where traditional machine learning requires a lot of text preparation such as removing stop words and case, Deep Learning models require minimal preparation of text and often do best when contextual text features are left intact.

The machine learning models in Deep Learning are "deep neural networks", which are neural networks having one or more "hidden layers" that mathematically model and connect relationships in the data - between the starting input and finishing output layers - and hence are referred to as "deep". Deep Learning is referred to as:

A sub-field within machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture. (Deng, 2014, p. 7)

Simply put, deep or hierarchical learning allows for more complex modelling of data by creating and combining many lower-level features. Deep Learning belongs in a domain called representation learning, where ML models automatically learn features or representations of the data. That is, the model creates features that *it* understands from relatively raw data and does not require features to be engineered by a system expert. In training a DL model, the feedback from how well the model is doing goes back to the model, which can act on the feedback to make its own adjustments to how it should deal with the data. This removes the burden of understanding and engineering the domain from a human being, and it can be more truly said that the *machine* is learning (Lecun et al., 2015).

Recently health-related research has embraced Deep Learning, as the benefits of a model that can rapidly learn how to make use of raw data given enough examples is eminently suited to many health-related data sets, initially in image processing such as radiological reports, and as NLP techniques in DL have improved, in social media text processing (Ravi et al., 2017). Although lexicon-based approaches have often been used to classify ADR mentions, they are not really suited for social media processing due to the effort of manually incorporating numerous non-medical terms, whereas DL models have demonstrated the ability to learn mappings between informal and medical terms (Chowdhury et al., 2018).

Deep learning typically requires a lot of labelled data, so provided it can be labelled, the large quantity of social media data available makes it a suitable candidate for processing with deep learning models. Although labelling is usually a manual process it can be automated to a large extent through a reinforcement ML process called active learning (Settles, 2012), and labelled text can be adapted to similar domains, which is known as domain adaptation (Daume III & Marcu, 2006). Deep learning models trained in one domain can also be adapted for use in another similar domain, even if the new domain does not have a lot of labelled data — this is called transfer learning.

Deep learning models

The default deep learning model is called a "fully connected" network — it consists of nodes in successive layers that are all connected to one another, and therefore all capable of contributing to the propagation of information throughout the network. While a fully connected network is capable of handling NLP tasks, learning through all those connections can lead to problems as the models get larger, and there are other architectures that are more suitable in practice.

A major complexity of learning how to handle text is due to the sequential nature of language, words are not discreet units of information, they have a context which can often be quite spread out in a text. To cope with this, *sequential* neural networks have been developed — they can preserve something of the long-term context of each word.

A sequential network calculates and preserves the relationships to context words that are important for each word, while discarding words that are not needed for a word's contextual meaning. Thus, the models are not forced to represent all the word relationships but only those that really matter while traversing the document. The basic sequential model is the Recurring Neural Network (RNN). With larger networks, standard RNNs can have problems of either losing or amplifying the influence of a specific input as the network cycles around the

connections, this is called the vanishing or exploding gradient problem (Graves et al., 2008). Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014) are more advanced RNN architectures which are designed to deal with this problem, by having extra connections to preserve long-term relationships and to forget unwanted connections. Over the last few years more powerful implementations of sequential networks called *attention* networks (Vaswani et al., 2017) have been developed. The current state-of-the-art in attention networks are called Transformers. These are developed by large organisations, for instance Google's Bidirectional Encoder Representations (BERT) (Devlin et al., 2018); Facebook's Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019); Google/CMU's XLNet (Z. Yang et al., 2019) and Facebook's XLM (Lample & Conneau, 2019). These are the types of models often used in transfer learning, described in the next section.

An alternative approach to developing a sequential understanding of a word's context is to map its relationships to all immediately surrounding words, akin to constructing n-grams from frequently associated words, which can be implemented via Convoluted Neural Networks (CNN). Originally developed to process image information, CNNs have proven to be very efficient in many NLP tasks (Kim, 2014). These are quite effective in handling shorter texts that tend to have key phrases or words in proximity, and do not require sequence modelling of the entire texts of a document. The advantage of CNNs is that they reduce the complexity of the network by translating inputs via a series of filters into *convolved* features that simplify the data the network needs to handle. They can perform reasonably well on smaller amounts of data and are also relatively quick to train, compared to the requirements of sequential architecture models (W. Yin et al., 2017).

Transfer Learning

Transfer learning with a deep learning-based framework uses DL models where the domain and the task used in training and application of the models can be different (Aggarwal & Zhai, 2012). It is a mean for transferring knowledge from an auxiliary domain to a similar but different target domain, and has been used in many real-world applications such as natural language processing, computer vision, biology, finance, and business management (Lu et al., 2015). With transfer learning for NLP, models are typically very large and are trained on a massive amount of text such as the entire Wikitext corpus, then these models can be fine-tuned to quickly learn the nuances of a new text dataset. The resources required for training these models are often immense and only obtainable to large organizations like Google, but the

trained models can be used by an ordinary ML practitioner with access to a GPU-equipped PC. Using transfer learning offloads the deep learning requirement for massive, labelled datasets and very large DL models to the organisation doing the training, the users of the models can then make do with much less data and less computing resources. Transfer learning has been in increasingly used to achieve state-of-the-art results in many NLP tasks. The top four systems designed for classification of tweets mentioning adverse drug reactions in Social Media Mining for Health (#SMM4H) shared task 2019 used Transformer-based transfer learning techniques (Weissenbacher et al., 2019).

Transformer models used by this study are based on Transformer language models, which utilize multi-headed encoder/decoder attention mechanisms, that dispense with recurrence and convolutions entirely (Vaswani et al., 2017). They use intentionally masked sections of text to learn to predict the most probable words in sentences, and use byte-pair-encoding (BPE) (Sennrich et al., 2016), which copes with unknown words by encoding them with subword units. RoBERTa ("Robustly Optimized BERT Pretraining Approach") was developed by Facebook to improve on Google's original BERT ("Bidirectional Encoder Representations from Transformers"). RoBERTa improves on BERT by removing its next-sentence pretraining objective; by using larger mini-batches and learning rates; and using an order of magnitude more data and for a longer time than BERT was trained on. RoBERTa Large was the largest model of the available RoBERTa models on the Hugging Face site.

In this subsection essential NLP techniques utilized in mining and processing health-related social media text were reviewed. Unsupervised techniques, particularly topic modelling, were described as very effective for discovering the main subjects in a text collection without requiring labelling. Supervised techniques were characterised as using labelled data to guide machine learning models for goals such as text classification. Deep learning and transfer learning were reviewed as the current most powerful machine learning techniques for natural language processing. As well as the labelling requirements of supervised models, it was noted that traditional approaches require more data set up and model tuning compared to deep learning approaches, which require a greater volume of data.

2.7 Chapter 2 summary

This chapter provided background and context for the research — starting with an explanation of the importance of vaccine surveillance as a vital component of ensuring vaccine safety and describing the deficiencies in traditional vaccine surveillance reporting systems, especially their propensity for delay. Research into the growing use of social media monitoring to obtain

timely health-related messages was evaluated, and the benefit of identifying real-time social media-derived Vaccine Adverse Event Mentions (VAEM) was explained.

This led to an examination of the text mining and Natural Language Processing (NLP) techniques that are relevant to identifying VAEM in social media posts. The chapter described the suitability of unsupervised machine learning clustering methods to automate the identification of subjects in large volumes of texts. Topic modelling was reviewed in depth as being an often used and highly effective clustering technique, and it was emphasized that appropriate evaluation measures are needed to derive the most suitable topic models. The highly effective VAEM filtering technique described in Chapter 5 uses unsupervised topic modelling with a small number of labelled records to enable evaluation of the most useful topic models for the filtering task.

The chapter next introduced Deep Learning models (deep neural networks) by highlighting their superiority to traditional machine learning models when working with large datasets like those found in social media. Transfer learning was described as a method to harness powerful pre-trained deep learning models for text classification tasks. In Chapter 6 a range of traditional and deep learning models are evaluated for their applicability for classifying potential VAEM-containing tweets, and it is seen that deep learning and in particular transfer learning are the most accurate classifiers provided they are given enough training data.

.

3 Research Design

3.1 Chapter overview

This chapter describes the research approach that was taken to explore the problem and solution domain that relates to identifying Vaccine Adverse Event Mentions (VAEM) in social media. The approach is described both in terms of research stages and in terms of a framework of the NLP techniques and processes that were used in the practical stages of the research. The framework provided a structure for developing a method to reliably extract VAEM from Twitter data, which was eventually formalised as the VAEM-Mine method (Section 3.5).

3.2 Research approach

The approach first consisted of a stage of domain exploration — conducting literature reviews to understand and describe the problem and solution domain. Twitter was chosen as data source and tweets were gathered and prepared for data exploration using topic modelling. The viability of Twitter data for VAEM detection was established and a pipeline of topic modelling and classification was developed, leading both to a dataset of VAEM and to a method for VAEM filtering that can be adapted to any similar task. Figure 2 summarizes these stages, as high-level descriptions of the approach that was taken when planning and conducting the research.

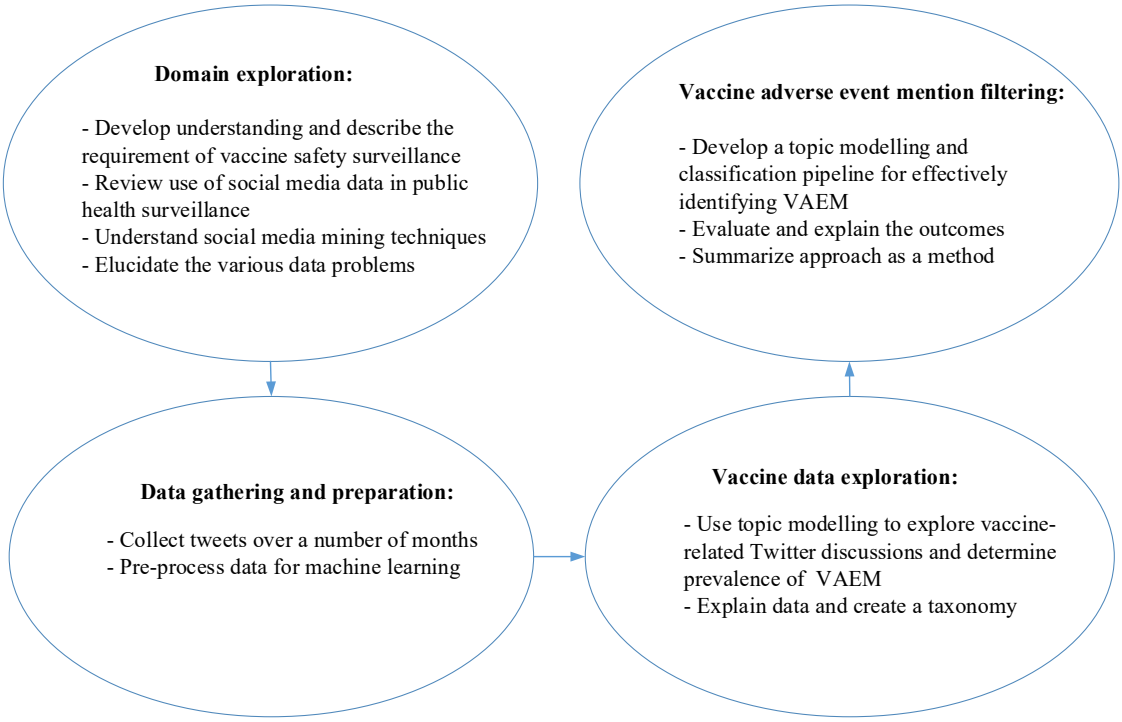


Figure 2: Research approach

Analysis of the Twitter data showed that tweets containing vaccine adverse event mentions tended to have a common structure, they were complaints about commonly experienced effects following a recent vaccination — physical effects like arm-pain and fever, very often with a reference to the cause, such as a "flu shot". There were very few mentions of severe reactions, and many of these when they were present were in discussions about a previously reported event or controversy, not in a current personal experience. Our domain expert confirmed that the numerous reports of commonly experienced effects following a recent vaccine that we were finding in Twitter posts was the kind of data that was needed, as it allowed for observation of trends. The expert also confirmed that this kind of data was not available through formal reporting systems, which focus mostly on severe events.

3.3 Research process

Therefore, the task of identifying VAEM was understood as requiring a process that could identify the *language* used to describe personal mentions of common health ill-effects, to discover the relevant *discussions*. The process was not required to identify specific adverse events, and it did not need to categorize the adverse events. Consequently, the practical goal of the research became one of isolating tweets that contained the target language, which can be characterized as vaccination-related "personal health mentions". This process is summarized in Table 2 as an overview of the *stages* of the process, with the motivations, critical aspects of the methods, and outcomes of each stage.

The *motivation* for the initial data collection stage is described as obtaining social media data *likely* to contain VAEM discussions, so that NLP techniques could then be used to identify those that contained VAEM, by targeting the language of the posts. This meant that the data collection *method* was to target a social media source that would be expected to have a high volume of posts describing common vaccine-related effects, but at the same time not to limit the search to only posts that had specific adverse mentions. Twitter was clearly a suitable data source for this, as tweets tend to be extemporaneous and declarative. Tweeting is used by many people to make a brief running commentary about often inconsequential personal experiences, just for the sake of having a general social media presence.

Topic modelling was the next stage of the process, as it was effective at gathering the VAEM tweets into very few or even one topic. Its *motivation* was to obtain a filtered subset of the data, so that classification had a simpler dataset to deal with, but topic modelling itself was not meant to act as a classifier. The distinctive feature of the topic modelling *method* was its evaluation technique. For this, a set of likely topics, including VAEM, were decided upon after manually

examining the data, and a set of 1,400 of the tweets were labelled with these topics. Topics were evaluated with consideration to how well they identified the labelled VAEM and assigned recall and precision scores based on their proportions of VAEM vs other labels. The most effective topics were those that brought together the most VAEM with the least of non-VAEM.

Table 2: Research tasks - Motivation, methods, and outcomes

	Data Collection **Chapter Four**	**Topic Modelling** **Chapter Five**	**Classification** **Chapter Six**
Motivation	- Obtain vaccine-related social media data that likely to contain VAEM discussions - Utilize the data to help evaluate NLP techniques for identifying the language characteristics of VAEM discussions	- To extract posts most like VAEM: i.e., those containing discussions around personal concerns and experiences with vaccines and of vaccine reactions… to simplify these subsequent VAEM identification tasks: • labelling datasets for classification • classification learning process - To understand the distribution of topics in vaccine-related posts	To train classifiers to identify VAEM as a binary choice in the now mostly personal vaccine-related discussions, which include VAEM • Use the data extracted from topic modelling to ensure classifiers are dealing mostly with VAEM and similar personal health mentions • Reduce the data complexity and simplifying the task to a binary classification to improve the models' capacity to detect VAEM
Method	- Targeted Twitter as a platform, due to its: • extemporaneous, brief, declarative posts • real-time nature • high volume • public availability - Used Twitter API with wide-ranging search terms, such as "vaccination" and "flu shot"	- Manually examined tweets to determine the major categories, with VAEM as a distinct category - Labelled some tweets with those categories, evenly apportioned - When training models, observed how the labelled tweets were distributed in the model's topics - Identified the topics that most clearly contain labelled VAEM tweets - Extracted the tweets from those topics, for use in the classification step - Evaluated the best trained topic model when applied to a separate dataset	- Assessed classifiers performance when trained on • medium sized vs larger datasets • imbalanced vs balanced datasets • and evaluated on imbalanced vs balanced test datasets - A range of models were tested: traditional classifiers, neural networks, and transfer learning with Transformers - Hyperparameter & vectorization settings were optimized - A rules-based model was created as a baseline; various extra features were also evaluated
Outcomes	- 811 thousand posts were collected over a year - Cleaning, including de-duplication, reduced this to 688 thousand posts - Examination of the data showed very few VAEM - Later analysis showed VAEM to be no more than 1.5% of the data	- 85.6% of data was eliminated after extracting only topics that were likely to contain VAEM - The extracted 14.4% of data consisted of around 99 thousand records, which were manually labelled, as VAEM and non-VAEM - Percentage of VAEM in extracted data was around 10% - A taxonomy of vaccine-related posts, based on the models' topics	- The best models classified VAEM with F1-Scores over 0.9 - 94% of all VAEM were identified via the combined effect of extracting VAEM via topic modelling, followed by classification over the extracted data - 98.6% of initially collected data was eliminated through these processes, mostly non-VAEM, with only 6% of VAEM also lost

The trained topic models needed to be tested against new data, so topic modelling was applied through two data collection stages, with the first lot of data being used to train the models, the second lot being used to evaluate them.

Although the topic modelling scoring mechanism borrowed concepts from classification measures, the topic modelling stage should be thought of as a filtering mechanism, and not as a classification process. That is, it was used to eliminate nearly all tweets that were not like vaccine-related personal health mentions, to simplify the task presented to the following classification stage.

The *motivation* of the classification stage was to determine which of the mostly personal health mentions obtained from the topic modelling were describing VAEM - and classification was therefore framed as a simple binary task. The aim was to find the greatest number of VAEM as accurately as possible, so that downstream consumers of the data could use the data to determine trends. Crucially, as the models were dealing with already filtered data, it was much easier to get higher performance from the classifiers than it would have been, had it been necessary to compensate for class imbalance using other data manipulation or classifier techniques, or if multi-class classifiers had been needed.

Classifiers used were well known and available to other researchers and were preferred if they did not require advanced expertise or special features to get the best out of them. For instance, the classifiers should learn what was distinctive about the language of VAEM based on the labelled examples only. If components of the language such as negations were important, then the classifiers should learn this based on how the texts were labelled, rather than requiring extra features added by the ML practitioner.

As the topic models used data that had been collected in two stages, the classification method also utilized the filtered outcome of the topic models following these two data collection stages. That is, models were trained and evaluated with the first lot of data, then were re-trained and evaluated with the combined data of both lots of collected data, with some data being set aside for testing purposes. This was instructive as it highlighted the increasing performance of the classifiers relative to training data, and although the positive effect of adding data was expected it was considered worthwhile to quantify this in the evaluation of the classifiers. The best classifiers proved to be Transformer models, once sufficient data was made available.

In the *Outcomes* sections, the table tracks the progress of the identifying and isolating VAEM through the stages of the processes — these were tangible outcomes, chiefly the posts containing VAEM, but the topic modelling stage also produced a taxonomy. Evaluations of

the percentages of VAEM in each stage help with understanding the effectiveness of the process as it proceeds through the stages. The conclusions that only 1.5% of the original data contained VAEM, and that the overall process discarded 98% of the irrelevant data to identify 94% of the VAEM, were made after the entire work had been completed - this is explained in Section 7.4. The technologies and processes outlined above can also be described as a framework, which is described next.

3.4 Framework

As presented in Figure 2 and Table 2, the research process started with domain exploration, then was followed by the practical stages of data gathering and preparation, data exploration and vaccine adverse event mention filtering, which required topic modelling and classification steps. Specific details of the process are presented in Figure 3 as a framework.

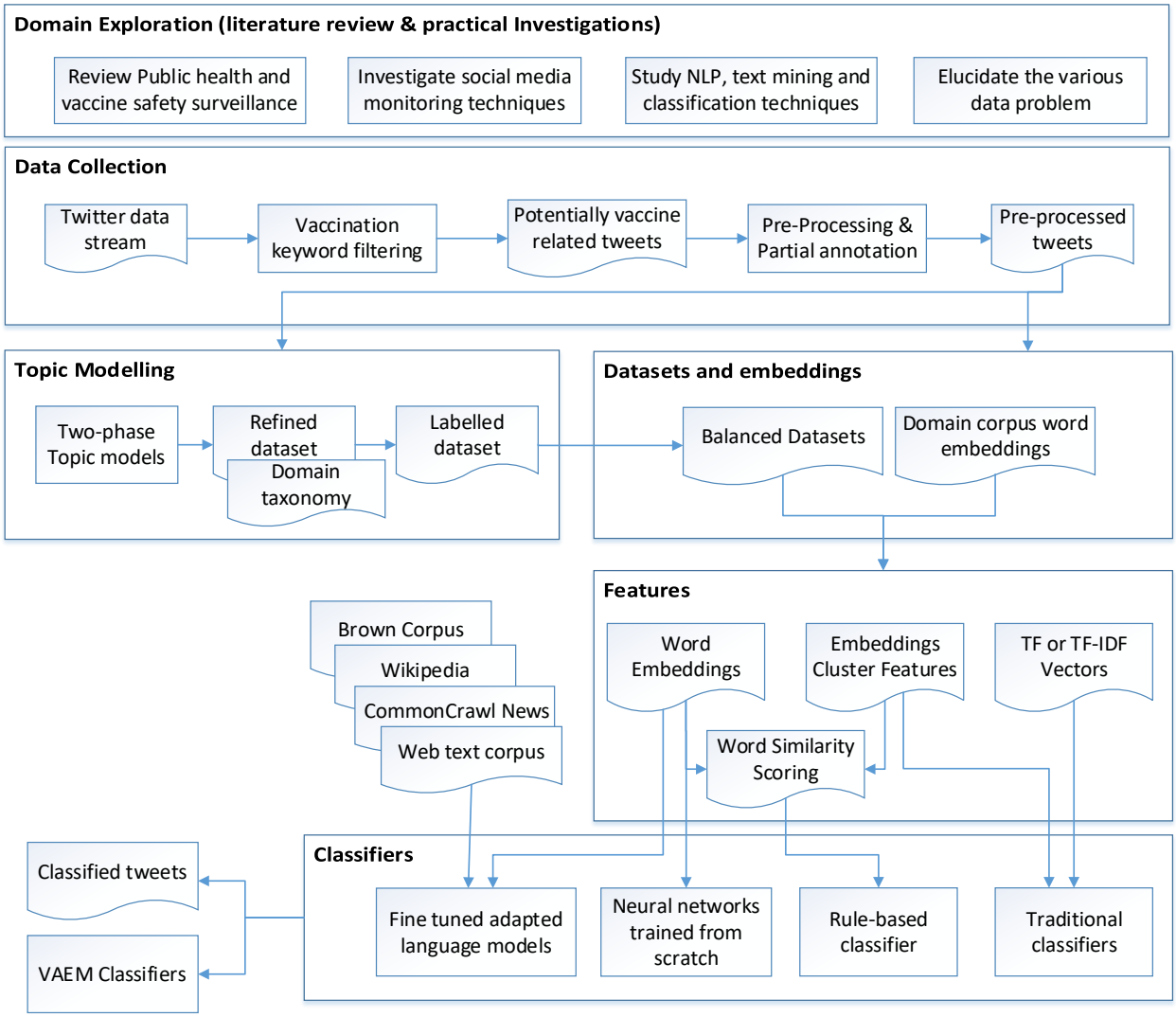


Figure 3: Research design framework

The next section describes the framework components in more detail.

3.4.1 Domain exploration

The first stage of the research was to conduct a literature review around the importance of vaccine safety and vaccine safety surveillance (as described in Chapter 2). The limitations of traditional data sources and established reporting systems were surveyed, followed by an investigation of the advantages and challenges of using social media data as an alternate data source. This included reviews of the use of social media in public health opinions and in surveillance, specifically disease surveillance and medication safety. The literature review highlighted the difference between social media monitoring for adverse drug reactions and adverse events following immunization. The review explained the point of this research, which is to monitor for the general level of vaccine adverse event mentions, rather than looking for specific severe adverse events. It was noted that this is more akin to social media monitoring for personal health mentions, but that very little vaccine-related research has been done in this area. The existing methods and algorithms commonly used in social media monitoring were summarized and their limitation and strengths were explored. These covered automatic classification methods, including topic modelling and deep learning methods.

3.4.2 Data collection

Table 3 summarizes the problems and the potential strategies that are relevant to identifying VAEM in the Twitter stream, and these are applicable to any similar challenge of finding a signal amidst an immense quantity of other indirectly related social media posts.

Table 3: Strategies for data related problems

Problem	Description	Strategy
Noisy Data	Vaccine-related messages are overwhelmingly not about personally experienced adverse events	• Filter at source • Filter afterwards • Unsupervised clustering
Weak Signal	Even putting aside obscuring noisy data there is just not much posted that contains VAEM	• Remove overrepresented data • Increase signal through collecting more data
Sampled Data	Streaming media APIs sample data based on filters and contractual agreements, and there are many social media outlets	• Gather data for a long period • Get data from multiple social media sources

The third column of the table summarises the kinds of solutions that are applicable to each problem. For instance, noisy data can be filtered at the source by the search filter used when downloading streaming data and can be further improved by applying filters afterwards to remove unwanted data. The iterative process of removing noise from already downloaded data suggests rules that can be applied to refine the source filtering strategy. Although the subsequent identification of keywords or word combinations may suggest more restrictive filters for the streaming process, care must be taken to avoid elimination of valid data, so the best strategy is a balance of a filtered search followed by further filtering afterwards - where more sophisticated data processing techniques can be applied than those available in a social media API.

The Twitter streaming API was used with a search term that would gather a broad range of vaccine-related data without capturing too much extraneous data: "vaccination, vaccinations, vaccine, vaccines, vax, vaxx, vaxine, vaccinated, vacinated, flushot, 'flu shot'". The aim of the search was to gather wide-ranging vaccine-related discussions, including those related to the commonly used terms for flu vaccines, to provide a variety of data for evaluating VAEM text mining techniques. The approach is explained in Section 3.3, the Research Process, and Section 4.1.1 contains a detailed discussion of the reasoning behind the approach. The search was limited to English language tweets, but no geo-location restrictions were applied.

Twitter data was gathered from 7th February to 7th June 2018, comprising 400,097 tweets. A further 3 months of Twitter data was also collected between 9th August and 12th November 2018, containing 401,482 tweets, and another set of 9,431 tweets were collected between 7th May and 20th July 2019. This resulted in total of 811,010 tweets, almost a year's worth of continuous downloading. Ethics approval for this study was granted by Monash University Human Research Ethics Committee (Project ID: 11767).

A manual examination of the data showed that few of the discussions contained adverse events mentions. The data was deduplicated, cleaned and prepared for topic modelling, with an aim to understand what the main topics were and to get an idea of the extent of VAEM in the data. The data was also processed into Word2Vec word embeddings, for later use in the planned classification stages, using Gensim Word2Vec and skip-grams. Detail of the data collection and preparation can be found in Chapter 4, "Data collection and preparation".

3.4.3 Two stage topic modelling

Topic modelling became the major contributor to understanding the Twitter data and developing an approach for extracting VAEM from the data. Topic modelling revealed the

main topics in the data, assisting with the development of a taxonomy of vaccine-related tweets, and revealed that VAEM were identifiable as a distinct topic. An evaluation approach was developed, which consisted of labelling a small number of the tweets and observing when topic models put tweets that had been labelled as VAEM into (ideally) one topic. The best model located almost 100% of the labelled VAEM into one topic, alongside similar tweets, which suggested that topic modelling could be used as a filtering mechanism. Further experiments showed that a second stage of topic modelling could be applied to filtered tweets, with a resulting VAEM-focussed dataset, which was used for developing the features that would be required for classification. Chapter 5 "Topic modelling" describes this.

3.4.4 Datasets and embeddings

The output from the topic modelling was labelled as either VAEM or non-VAEM and then, after conducting training experiments with imbalanced data, the data was balanced by removing excess non-VAEM to contain roughly equal amounts of each label. These were then used to create the training, validation and test datasets used in the Classification stage. Additionally, Word2Vec embeddings (Mikolov, Chen, et al., 2013) were created from the pre-processed tweets — that is, from the entire cleaned data rather than just from the labelled tweets. The embeddings were used by the neural networks as word vectors and were also used to develop additional features.

3.4.5 Features development

TF or TF-IDF vectors were created for use by the traditional classifiers. Apart from their standard use by neural networks, word embeddings were also used to create similarity scores that were utilized by a rule-based classifier, and for clusters that were also evaluated with traditional classifiers. Appendix D contains detailed descriptions of these. Various additional features were experimented with, see Appendix F, "Feature engineering results".

3.4.6 VAEM classification

Although topic modelling on its own was not able to isolate VAEM with much precision it was highly successful in gathering VAEM into topics that included other similar messages and excluded most of anything else. Text classification models were then used to further isolate the vaccine adverse event mentions. Benchmarks for the classification models were created by assessing traditional machine learning approaches (S. Wang & Manning, 2012). To understand their success more clearly a benchmark of a manual rule-based classification process was

developed. Ensembles of models were also evaluated - ensembles frequently outperform individual models because each model deals with features differently (Heaton, 2016). Various neural network-based classifiers were also assessed, ranging from Convoluted Neural Networks (CNN), which suit detecting key phrases and patterns that do not require extensive language interpretation; to Recurring Neural Networks such as Long Short-Term Memory models (LSTM) that are designed for handling language data; and culminating in the current the state-of-the-art Transformer language models, which were trained on various large text corpuses. Chapter 6, the "Classification" chapter of the thesis, deals with this.

3.5 The VAEM-Mine method

The application of the various processes described in the framework can also be described as a *method* - that is, as the steps or pipeline required to obtain a classified set of tweets. Figure 4 illustrates the VAEM-Mine method. 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 refined to extract the valued VAEM-containing data. The VAEM-Mine method consists firstly of processes of data collection and data preparation, then as topic modelling and classification phases, both of which first require models to be trained before eventual deployment into a working pipeline.

The method should be understood as a map for the choices of either training the topic models and classifiers, then having trained them - of using the deployed models to filter and classify new incoming tweets. 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 then the work of training the topic models begins. For training topic models, the first step is to label some examples of the subject of interest so that topic modelling scoring can be applied, to enable assessment of how well topic models perform in the task of filtering the data to contain the label of interest in just one topic, or in just a few topics. Labelling the records and scoring are only required in the training phase and are not part of the eventual deployed pipeline, and a large quantity of records is not required – around 150 to 200 of each of the potential major topics, as judged by manual observation of the tweets. This is explained in Section 5.1.3. 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 will identify 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.

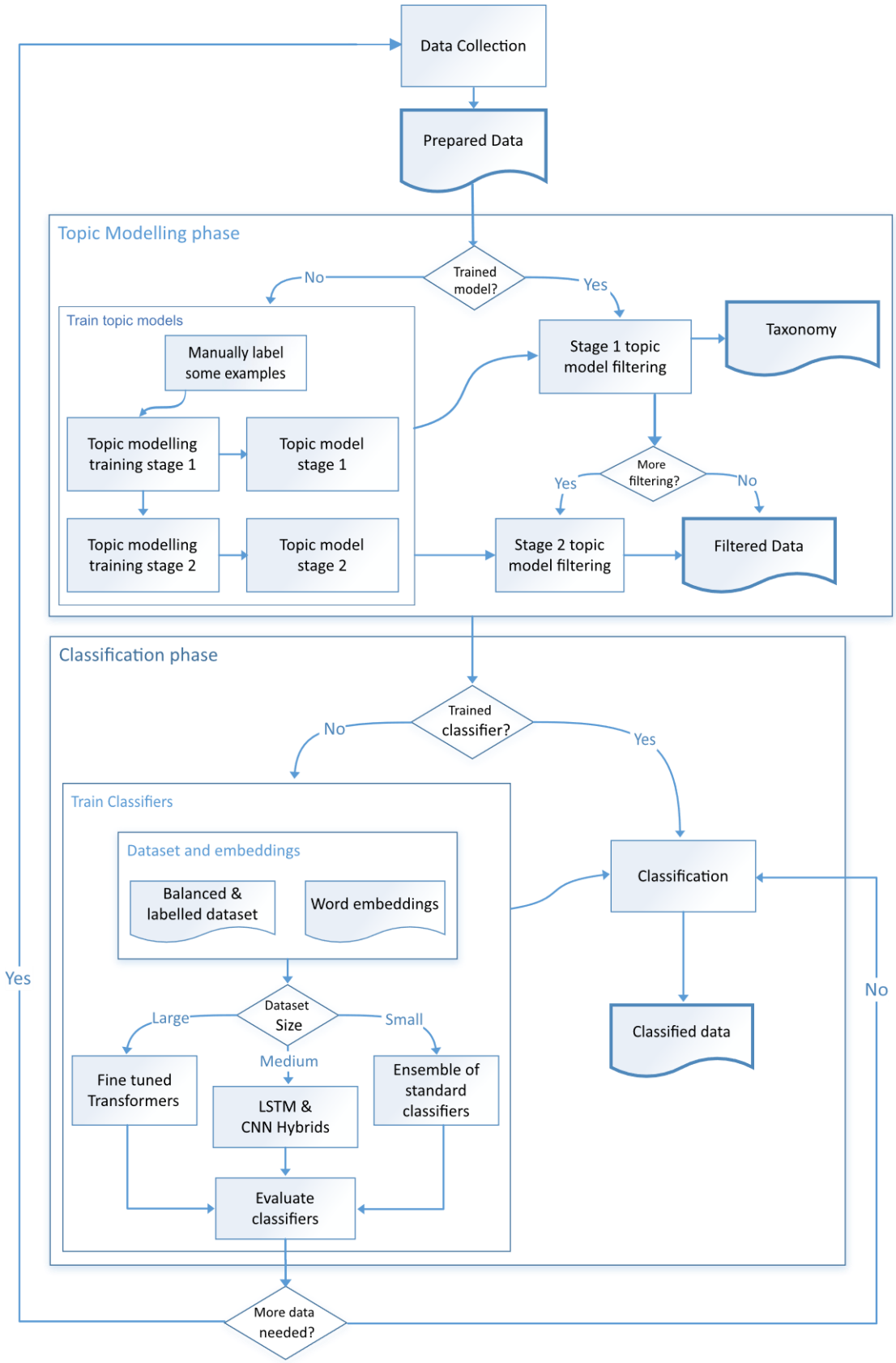


Figure 4: VAEM-Mine Method

Having trained the topic 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 of the first-stage topic model, or 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 then passed into the classification phase. Entering the classification phase for the first time, the answer is No as to whether a trained classifier exists. Consequently, 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. Dataset size is a factor. Is there enough data to fine-tune language model-based classifiers such as Transformers, or would less demanding neural networks such as CNNs or LSTMs, or even traditional classifiers, be more appropriate? These questions are most likely only answerable after assessing a range of classifiers and comparing their relative performance. If the scores obtained are not meeting expectations for the classifiers — for instance, an F1-Score of 0.8 or more — a decision may be made to collect more data and to repeat the process.

In this research, it was found that the initial labelled dataset of 3.5 thousand records was insufficient to train Transformer models, but that hybrid CNN and bi-directional sequence models (e.g., a bidirectional GRU) were the best performers with F1-Scores around 0.8. With further data collection resulting in a dataset of 20 thousand labelled records all the models benefitted, but the outstanding result was an F1-Score of 0.9 obtained from the Transformers, which surpassed every other model. This is described in detail in Chapter 6 “Classification”, and a comparative analysis can be found in Chapter 7 “Evaluation”.

Once the classification training is complete then the method is ready as an end-to-end pipeline to be applied to any new incoming data. The crucial mechanism of the method is the pipeline of steps that are taken to increasingly refine the data for the extraction of VAEM, which requires the applications of trained topic models and classifiers in a sequence. However, these could be used separately. That is, the topic models can be used as a filter for raw data to create a smaller and homogeneous dataset for the classifiers, or their output can be used as-is, without classification. The pipeline specifies that training the classifiers requires assessment of all potential models against the available data, and that classifier choices and performance depends on how much data is available. Therefore, the best possible result may require the collection of more data, re-training and re-evaluation, until the pipeline is ready.

3.6 Chapter 3 summary

Chapter 3 introduced the research approach, firstly on a high level as stages of literature review, then data gathering and data analysis, followed by the application of techniques to identify and obtain Vaccine Adverse Event Mentions (VAEM) from social media. A description of a framework of NLP techniques and processes was used to make a more detailed assessment of the various components of the research, which included references to where these are explored in depth in the thesis. This was further explained in terms of the VAEM-Mine method.

4 Data collection and preparation

4.1 Chapter overview

Data collection and pre-processing are preliminary steps of the proposed framework, the interface between the data "at large" and the techniques that need to be applied to the data to discover meaningful insights. The research is predicated on discovering consistent VAEM-related information embedded in extremely diverse and mostly unrelated volumes of social media data. The data collection aim is to obtain as much relevant data as possible without excluding too much by using overly specific search terms.

Data pre-processing is dedicated to cleaning and pre-processing that data, then preparing it for use by the various downstream NLP technologies that are employed. The next section summarizes the general data processing techniques used in this research. The following topic modelling and classification sections each contain more specific explanations of their respective data preparation requirements, and details of the datasets. Those sections refer to processes performed in the topic modelling stage for filtering data to obtain more VAEM-like data subsets, so a basic understanding of the topic modelling processes is required. The classification section also refers to classification processes. Summary descriptions of those processes are presented with the dataset discussions, together with references should a more complete understanding be desired by the reader.

4.1.1 Social media (Twitter) data collection

The short, 280-word, message length of tweets was thought to be an advantage for this study, since it was likely that only one subject would be found per tweet. Another important feature of Twitter engagement is that it is used for making casual, extemporaneous, personal messages, which is the type of data the research needed to find.

Twitter data was initially gathered from 7th February to 7th June 2018, comprising 400,097 tweets, which covered the North American flu season, which amounted to a daily average of 3,390 tweets. A further 3 months of Twitter data was collected between 9th August and 12th November 2018, containing 401,482 posts, and combined with another set of 9,431 tweets collected between 7th May and 20th July 2019. The data from the first data collection round was used for training topic models and for initial testing of classifiers. The data from the second data collection round was used for evaluating the trained topic models and for further testing of classifiers.

The Twitter Streaming API was used with a search for "vaccination, vaccinations, vaccine, vaccines, vax, vaxx, vaxine, vaccinated, vacinated, flushot, 'flu shot'". Specific vaccines and disease-related terms were not targeted, as analysis of the Reddit data (*Reddit's publicly available comment dataset*, 2015) (detail in Appendix I) had shown that many of the terms are ambiguous or polysemic and introduce too much noise. For instance, although the term MMR (meaning the Measles, Mumps, Rubella vaccine) might be infrequently used in social media discussions of personally experienced AEFI, the term is extensively used by the online gaming community where it means "Matchmaking Rating". The terms "shot" and "jab" are associated with receiving injections but they are also used in too many other contexts to be useful search terms, whereas "flu shot" as a phrase was useful. Furthermore, the domain expert had advised that the strongest and most relevant adverse event mentions would likely come from flu vaccine-related incidents, due to the fluctuating nature of the seasonal flu vaccine compared with the much more stable nature of childhood vaccines.

Arguably, further specific search terms could have been included to increase the likelihood of finding adverse event mentions. For instance, expanding the "flu shot" query with combinations like "flu jab", "influenza injection"; and using terms such as "AEFI" or "reaction following immunization". However, the goal of the research was to determine how to find VAEM based on the language in the texts, not to elicit the best search terms for catching specific instances of vaccine adverse mentions — so the author decided not to add additional terms to the most often used expressions such as "flu shot". Initially, some of these specific terms were trialled, but it was found that posts that featured VAEM, other than those from a flu shot, almost always appeared in combination with a general term such as "vaccination". Furthermore, technical, and formal terms were hardly ever used to describe VAEM. Therefore, given the research goal, it was decided to just use the general terms. The limitations of not using more specific terms are discussed in Chapter 8.3.

Most posts were not reporting VAEM, and even those that might contain VAEM could turn out to be something different. Table 4 illustrates a sample of VAEM-related tweets — almost all of them contain personal anecdotes with phrases such as "I got" and "my arm"; indications of pain associated with words like "vaccinated"; and include words such as "today" and "yesterday" indicating a recent event. Although examples 5, 7 and 8 are not VAEM, the type of language used in all these examples is consistent and unlike other vaccine discussions, which lack the combined elements seen here. Note that example 7, which relates to a possible vaccine reaction in a puppy, would not be labelled as VAEM for the classification step, even without

the qualification indicating that there was no reaction, as it was decided to consign all animal-related tweets to the non-VAEM category.

Table 4: Sample of language used in vaccine-related tweets

1	I got my second meningitis vaccine and it literally hurts so bad just to put on my backpack, it's a sign I shouldn't go to school for the rest of the week
2	I finally got the flu shot yesterday and now my body feels weird. I don't know if I'm having a reaction or what.
3	i spent all day thinkin i was a baby for having a sore arm after getting vaccinated. anyways turns out im like. rlly allergic to the meningitis vaccine i got :)
4	This flu shot got my armpit on fire
5	I'm proof of this! I got the flu shot and instead of feeling like I was dying for 5-7 days (actually diagnosed with flu on Friday), I was only achy, fever-y, and stuffy for three days in the middle. 6 days after starting the symptoms, I'm coughing, and that's about it.
6	I got the flu shot today and in that area it's fucking hurting .. is that normal ? Am I dying?????
7	So this dude brought his dog to the hospital today for some vaccines right, puppy passed tf out!!! We thought it was an allergic reaction, turns out, mf ain't feed the dog in 12 hours. The pup is 3 months. Smh.
8	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.

4.2 Data pre-processing

Texts need to be regularized through pre-processing so that statistical techniques can be applied to recognize language patterns, that allow for subject grouping. This section describes the pre-processing undertaken with this data to prepare it for topic modelling.

4.2.1 Removing unwanted tokens

Stop words

When looking for meaningful words to count and analyse very often there are words that have no particular significance - they may be words that are just used everywhere (“a”, “the”, “in”, “my” etc.) or words that are hardly ever used. The commonly used words are known as “stop words”. When assessing the impact of stop-word removal it was found that although the initial topic modelling benefitted from their removal generally the classifier models did worse with stop words removed – likely due to the important language clues contained in the intact text.

Punctuation, Numbers, Symbols

Punctuation and numbers are components of text that can also be removed to good effect, particularly with a bag-of-words approach, as measuring the incidence of punctuation mark and number usage is not normally required. Punctuation can however be utilised in more sophisticated NLP techniques to determine sentence termination, possessive nouns etc. Punctuation and numbers were removed. Other symbols such as emoticons are on the other hand used constantly to convey meaning, so should be retained. To regularize emoticons, they were converted into equivalent English words (e.g., " :(" was changed into "sad").

Non-predictive tokens

Special phrases such as URLs, hash-tags, Twitter names and retweet symbols were also removed. These phrases are not only repetitive and non-essential features that confuse the analysis of word significance but also are highly specific but transitory. For instance, a particular email address or signature phrase could be identified as being more significant than the words coming from the general discussion of a text, simply because it appears repetitively but not because it is part of the subject that the text discusses.

4.2.2 Pre-processing and adding features

Case and spelling

Words may assume different forms but have the same meaning - for instance "His" at the beginning of a sentence probably has the same meaning as "his" in the middle of a sentence, and a mixed-case typo "hIs". These probably do not want to be identified as different words, and so a common strategy is to make all words lower-case so that the vocabulary is trimmed to just one symbol per word, "his" in this example. In a similar vein spelling mistakes can introduce variations which could be corrected if desired. On the other hand, there may be times when case matters, texts dealing with the subject of God for example may deliberately use the word "His" (or maybe "Her"?) when referring to the deity and this carries a significance compared to an ordinary "his", likewise upper-case "HIS" may very likely be an acronym or an organization name. It was found that using lower case has benefited the topic modelling, but a copy of the original text was retained for later re-evaluation with the classifiers. Spelling mistakes were not corrected as the text is full of jargon and peculiar words, and most of them are deliberate - they get ignored anyway by the models, and the focus was on the significant words which mostly are correctly spelled.

Text normalisation - Stemming or Lemmatization

The various derived forms of words can be usefully reduced to a single form - either through stemming which creates a root form that might not actually be a real word, or lemmatization which uses a real word for the root. This is going further than the simplification of enforcing lower-case and removing punctuation, it is reducing many words into one word. For instance, "playing", "plays", "played" might all be reduced to "play" (though some algorithms might leave "playing" as it is); "am", "are", "his", "her" and "we" can be reduced to "be". Other studies have shown that lemmatization improves topic coherence (Martin & Johnson, 2015) and that lemmatization is preferred over stemming because it considers context (Mehta, 2020; Win & Aung, 2018). Lemmatizing benefited the topic modelling process, as topic modelling needed to reduce text variation in a large quantity of text to discover underlying similarities, and lemmatization was effective for this – see Section 4.3. The classification models did not do well with lemmatized text – Sections 6.3 and 0 describe the preparation and evaluation of lemmatized text with classifiers.

N-grams

As described in Section 2.5.2, n-grams are used to preserve phrases in a bag-of-words approach. In our work we have used one (i.e., single words), 2 and 3 grams, with or without stop words removal, and with or without the preservation of word-case. Topic modelling used n-grams created from the lemmatized corpus; some of the standard classification models benefitted from using n-grams, this was evaluated for each model.

4.3 Topic modelling data preparation

Data preparation for topic modelling consisted of a data cleaning process to remove obviously invalid tweets and duplicates before further processing. Cleaning initially consisted of converting from Unicode to plain text, eliminating URLs, converting to lower case, removing the retweet tag and @user references and the hash symbol from hashtags. Text-based emoticons were replaced with plain English equivalents. Documents having less than five words, or with a high number of non-unique (therefore repeated) words and documents with a low number of English words were removed.

Contractions were expanded prior to tokenisation (e.g., "Don't" was converted to "do not"); stop words were removed using the NLTK (a Python library for text processing) (*NLTK, Natural Language Toolkit*, 2018) stop words list, apart from the strongly indicative words "do"

and “not” (as in “do not vaccinate”); and bigrams and trigrams were created. The Gensim utility function *simple_preprocess* was used to tokenize the data, as it had a low overhead compared to spaCy’s (Honnibal, 2017) tokenizing approach, but spaCy’s NLP library was subsequently used to lemmatize, retaining nouns, adjectives, verbs and adverbs.

A dictionary of lemmatized terms was constructed using Gensim’s *corpora.Dictionary* function, and the dictionary’s document to bag-of-words (*doc2bow*) function was used to assemble a document corpus. After some experimentation it was found that trimming the dictionary had a beneficial effect on performance and coherence, so words that occurred in less than 20 documents or in more than 50% of the documents were removed from the dictionary and subsequent corpus. After trimming, the token count was reduced by 90% (e.g., from 100,148 to 9,986 in the first dataset), and the time to construct a Gensim LDA model was reduced by 80% (e.g., from one hour to ten minutes), and topic coherence increased by 10%. As a result of these steps some documents had been stripped of all their words, and therefore were removed.

4.4 Topic modelling datasets

The initial dataset used for topic modelling consisted of 400,097 tweets that were collected in the first 6 months of the data collection process. After the de-duplication step this was reduced to 341,507 documents, and data cleaning further reduced the data to 329,842 documents. After removing any empty documents resulting from the lemmatization step the document count was 328,822. This dataset was used to train the topic models. Part of the topic model scoring technique required a small number of labelled tweets, so 1,400 of them were labelled, following the guidance of the domain expert — see section 5.1.3 for detail.

Topic modelling used two stages, and the best second-stage topic model (*the DMM 9-topic model, see Section 5.6.1*) provided an opportunity to extract a VAEM-like subset of tweets, by taking the top 3 topics that concentrated VAEM — being topics 8, 9 and 1. The result was a extracted subset of 18,801 tweets, which retained the topic numbers and the original 1,400 topic labels. Additionally, labelling for VAEM (label 0) and Discussions (label 1) was completed for all tweets in the (best) Topic 8. This data was then used in the initial classification assessment.

An additional 401,482 tweets were collected in the second 6 months of data collection, as two datasets. The new datasets were combined and then compared with the earlier dataset. There were 15,740 tweets that were duplicates of tweets present in the earlier dataset, so they were removed. After the pre-processing and cleaning required for topic modelling there were

359,535 posts. These were processed by the topic models that had been trained on the first lot of data. That is, the 14-topic DMM model, previously ranked as the best model in *Phase One* of topic modelling, was applied to the data - and the 80,372 records that the model put into the best VAEM topic of the model (Topic 13 of the first-stage 14-topic DMM model — see Section 5.3.4) were retained as the most likely to contain VAEM. This data was combined with the previously extracted subset to re-train classifiers in a second classification round.

4.5 Classification datasets

The classification datasets were assembled from the tweets identified by the topic modelling process as most likely to be VAEM, and classification experiments were conducted in *two phases* — see Section 6.5 and Section 6.8.

Phase One: The results of the first round of data collection and subsequent topic modelling and extracting the most likely VAEM texts was a dataset of 18,801 — but after fully labelling and balancing it to contain an even number of VAEM and non-VAEM examples, the dataset size was a little over 4,100 records. This was used in the first phase of classification, which tested traditional classifiers and deep neural networks trained from scratch.

Phase Two: Data processed during the second round of data collection and topic modelling was added to the existing Phase One dataset to produce a combined Phase Two dataset of around 20,600 records. All the models were retested on the Phase Two dataset, and having more data allowed for additional deep learning models to be evaluated. The performance of all classifiers improved, and particularly that of the deep learning models. Sections 4.6 and 4.7 describe the datasets in detail.

4.6 Phase One classification data

The data extracted from the topic modelling phase was mostly like VAEM. The goal of classification was to classify that data as either VAEM or not. Therefore, all the data that had been exported from the topic modelling was binary labelled for classification, as either "VAEM" or "non-VAEM". These were manually decided by the author while observing all tweets, following the guidelines supplied by the domain expert. The task was reasonably straightforward due to the simple criteria for determining VAEM, which was to take any reference to an adverse effect in relation to a recent vaccination. There were only a handful of cases that required confirmation by the domain expert — see also Section 4.7.1 for examples of records that required expert judgement.

Phase One dataset creation took the output of the initial two-stage topic modelling — see Section 5.5. It consisted of the top 3 topics (topics 8, 9 and 1) of the best second-stage topic model (the DMM 9-topic model, see Section 5.6.1). Tweets containing references to the rock group "The Vaccines" were eliminated (Section 5.6.2), leaving 18,519 labelled documents. The DMM 9-topic model's topic labels and the original 1,400 manual labels were retained as fields in the data. Labelling for VAEM (label 0) and Discussions (label 1) was additionally completed for all tweets in the (best) Topic 8. This enabled the data to be assessed by topic number and topic model label when evaluating classifiers' performance while training with increasingly imbalanced data. This is described in Section 4.6.1.

Eventually a balanced dataset was assembled to complete the Phase One training, described in Section 4.6.2. A test dataset was extracted from this, which was utilized throughout all the experiments to allow a fair comparison of classifiers' capabilities (Section 4.6.3). The final dataset obtained after these steps is described in Section 4.6.4.

4.6.1 Experimenting with imbalanced datasets

The data that had been extracted from topic modelling including the topics numbers as a field. In Phase One, these were the "top topics" of the topic modelling second stage DMM 9-topic model, numbered 8, 9, and 1. Because each topic had increasing degrees of imbalanced data, they provided a simple mechanism for dividing the data into classification datasets of various degrees of imbalance. The 3 topics were used to create 4 datasets for classification, named "Best", "Combined", "Top Two Combined" and "All Combined".

The "Best" dataset contained just the records from the best topic, Topic 8. The "Combined" dataset contained all of Topic 8 and just the VAEM and labelled Discussions records (original topic model label 1) from Topic 9. The "Top Two Combined" dataset contained all of the records from topics 8 and 9, and the "All Combined" dataset contained all data from the three topics.

In Table 5 the topic numbers and topic model labelling information are included to illustrate the makeup of the four datasets.

Table 5: Distribution of original topic model labels in classification datasets

Label	Topic 8	Topic 9	Topic 1	Label Totals
Label 0 - VAEM	1,304	161	275	1,740
Label 1 - Discussions	1,697	64	754	2,515
All Others	48	5,365	8,851	14,264
Topic Totals	3,049	5,590	9,880	**18,519**
Datasets				**VAEM %**
Best Dataset	3,049			43%
Combined	3,549	(3,049 + 161 + 275 + 64)		49%
Top Two Combined	8,639	(3,049 + 5,590)		17%
All Combined	18,519	(3,049 + 5,590 + 9,880)		9%

The labels show that Topic 8 is largely either VAEM or Discussions, as identified by the retained topic model labels; and that included into the Combined dataset there were 161 VAEM and 64 Discussions from Topic 9, and 275 VAEM from Topic 1. That is, the Combined set had all the VAEM, but apart from the 64 extra discussions, without the addition of any other non-VAEM from topics 9 and 1.

Figure 5 shows the relative percentages of the VAEM vs non-VAEM labels in the four datasets - the "Combined" dataset is the most balanced with 49% of the data being vaccine adverse event mentions. The "All Combined" dataset is the most imbalanced with just 9% of VAEM present.

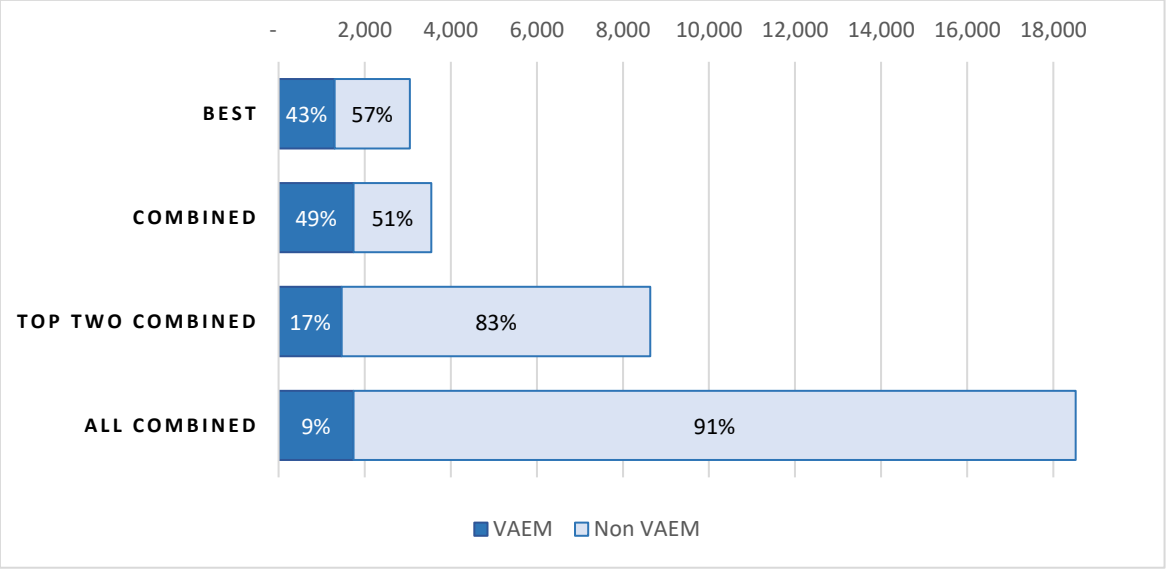


Figure 5: Proportion of VAEM in classification datasets

Section 6.5 contains an assessment of how the traditional models coped with these datasets.

4.6.2 Creating a balanced dataset

After completing the preliminary analysis of classifiers trained on varying degrees of imbalanced data, a final dataset was created that included all 1,740 VAEM labelled records from the three topics, all other data from Topic 8, and the Discussion (topic model label 1) records from topic 9. That is, a single dataset of 3,549 records, consisting of 1,740 VAEM records and 1,809 non-VAEM records - a ratio of 49% - see Table 6.

Table 6: Distribution of labels and topics in final Combined dataset

Label	Topic 8	Topic 9	Topic 1	Total	VAEM %
Label 0 - VAEM	1,304	161	275	**1,740**	**49%**
Label 1 - Discussions	1,697	64		1,761	
All Others	48			48	
Topic Totals	3,049	225	275	**3,549**	

4.6.3 Imbalanced (Victorian) test dataset

As part of the evaluation process, a test dataset was created for an investigation into local seasonal VAEM trends. The records were identified by looking through all the data in the geographical-related Twitter fields including UserLocation, UTCoffset, TimeZone and Place fields, to identify mentions of Victoria Australia and Victorian cities and towns, between 7th February and 7th June 2018. This period included the time when people were getting flu vaccines and the early Australian flu season of 2018. Based on the tweet text, the resulting 3,014 tweets were labelled as containing VAEM tweets or not, resulting in 93 VAEM records and 2,921 non VAEM tweets.

As with the other datasets the tweets were processed through the DMM topic model to eliminate most of the unwanted tweets. From 3,014 tweets with 93 VAEM the subset that was extracted contained 90 VAEM and 524 non VAEM — 614 in total, roughly a fifth of the original Victorian data. The resulting texts were assessed using the tweets' dates, and the trend for VAEM discussions was found to follow the seasonal application of flu vaccinations, see Appendix H.

4.6.4 Final Phase One datasets

The 614-record imbalanced test dataset consisting of Victorian tweets had some records that were in the existing final dataset, so these had to be removed from the final dataset to ensure the test data was unique. This reduced the number of records in the final dataset by 40 to 3,519

— consisting of 1,722 VAEM and 1,797 non-VAEM — a ratio of 48.9%. The final dataset and test dataset for Phase One are shown in Table 7. The traditional classifiers were freshly assessed using these datasets and the preliminary training of the neural networks also used them, and as more data was added (see Section 4.7.1) the imbalanced test dataset continued to be used as a benchmark when comparing classifiers performance.

Table 7: Phase One datasets

Label	Main dataset		Test dataset	
Label 0 - VAEM	1,722	48.9%	90	14.7%
Label 1 - Non VAEM	1,797		524	
Dataset Totals	**3,519**		**614**	

4.7 Phase Two classification data

Although it was hoped that all models could be assessed on the smaller dataset from the first phase of data collection, it became apparent that there was not enough training data to properly assess the neural networks. Deep learning requires a lot of examples for the models to learn well, and when evaluating the neural networks on the Phase One dataset their performance was only a little better than the traditional classifiers, and in particular the sequence-based networks such as LSTM and Transformer models did not perform as expected. Therefore, further evaluation of the Transformer models was deferred, and all models were re-tested, when the expanded Phase Two dataset was introduced.

4.7.1 Additional data collection

As previously described in Section 4.4, the initial data collection comprised 400,097 tweets, which was reduced via the two-stage topic modelling and data preparation for classification to 3,519 records, with a further 614 test records — a total of 4,133 records. Also described in Section 4.4, additional data collection and processing through the 14-topic DMM model, previously ranked as the best model in *Phase One* of topic modelling, resulted in an additional data volume of 80,372. The second stage 9-topic DMM model had also been applied to this data, but no further filtering was carried out. Instead, all the records were retained, along with their second-phase topic number.

These records were all manually labelled by the author as containing a VAEM or not, following the domain expert's guidelines. The domain expert verified 97 edge cases that needed checking for the author's judgement of VAEM, and there were only 10 that were corrected, mostly towards their being VAEM. For instance, "*I got the flu shot today and idk how I feel*" was confirmed by the expert as a VAEM, despite its vagueness, and "*Last time I took my flu shot, my balls was sore af*" was also tagged as VAEM, despite the sense of the event not having taken place recently. "*I always feel the most crappy the day after getting a flu shot*" was initially in doubt due to uncertainty about whether the user was referring to an actual recent vaccination. However, it was confirmed as VAEM as quite a number of tweets exhibited a similar structure and were certainly VAEM - e.g.: "*Every time I get the flu shot I get hella sick , when I say hella I mean hella. And I got the flu shot today and I already feel hella hot and hella light headed*".

Labelling all the data enabled verifying the previous conclusions about the topics most likely to contain VAEM, and also meant that all vaccine adverse event mentions were obtained. A balanced set of 16,251 records was extracted from the second phase data and combined with an additional 307 records that had been separately identified, resulting in a 16,558-record additional dataset. When combined with the Phase One data the total number of records was 20,691.

The Phase Two data was subsequently split into 15,730 records to be used for training and validation, and 828 records were set aside as a new balanced test dataset. These were combined with the Phase One 3,519 training and 614 test datasets, to provide 19,249 records for training and validation in Phase Two, and 1,442 records for testing. Table 8 shows these numbers.

Table 8: 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**

4.7.2 Balancing the Phase Two data

When balancing the Phase Two datasets all the VAEM were retained and the non-VAEM were under-sampled, which is one of the most popular and effective techniques for solving imbalanced classification problems (Kaur et al., 2019). This was appropriate for the dataset as there were plenty of the minority VAEM class. The final balanced counts were 9,995 VAEM and 10,082 non-VAEM, for a total of 20,077, excluding the imbalanced test dataset of 614 records (the total is 20,691 if they are counted). The topic numbers of the second-stage topic model that were available in the dataset were accounted for when extracting non-VAEM from it, so the data had a similar ratio of non-VAEM to VAEM records *per topic*. Figure 6 shows the balanced distribution of VAEM to non-VAEM labelled records per topic, the "Previous" data are the 3,519 records from stage 1 plus the 307 added records. The figure excludes the 614 records of the imbalanced test dataset.

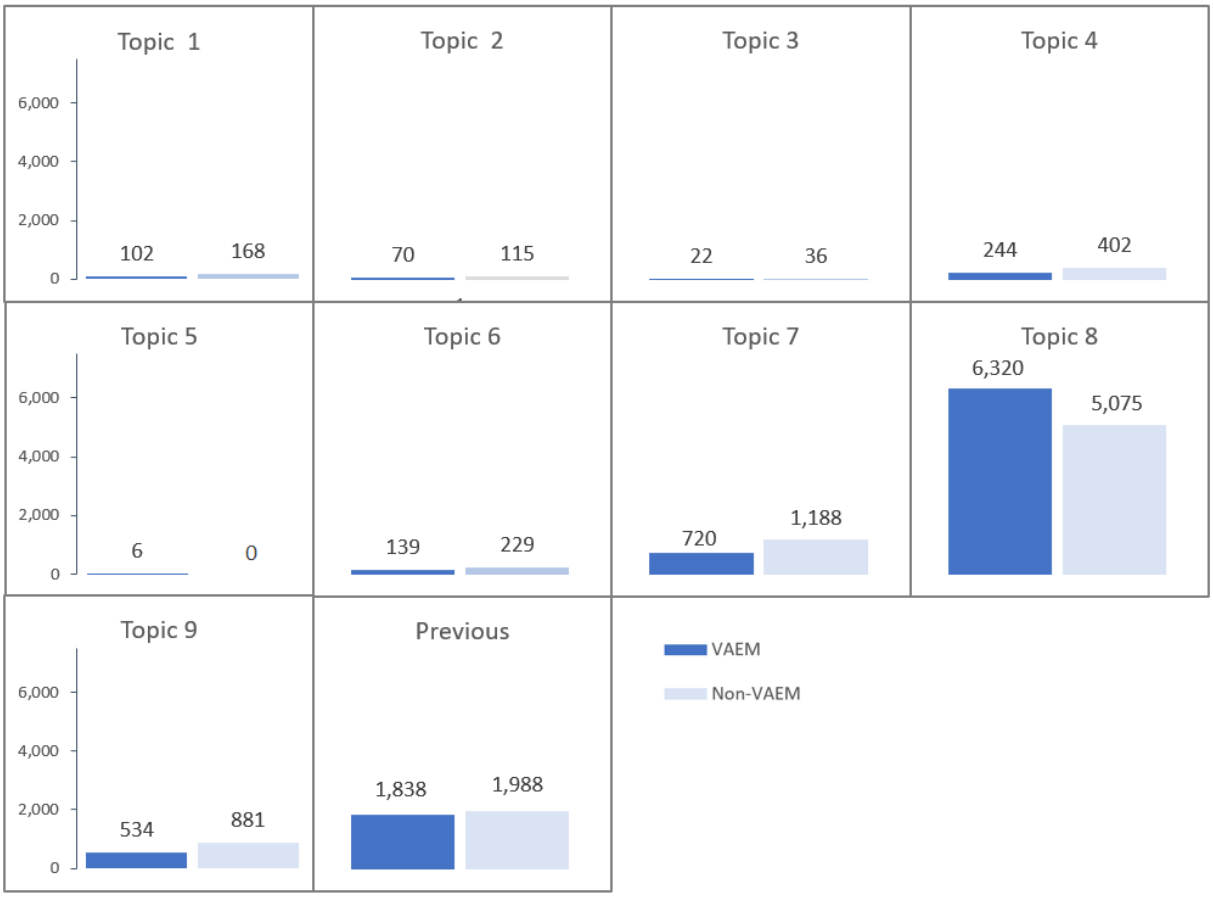


Figure 6: Distributions of balanced labels per topic — final combined datasets

4.7.3 Final Phase Two datasets

As described in Section 4.6.3 an imbalanced "Victorian" test set of 614 records was created during the first phase. It consisted of a set of tweets collected over four months having a

geographical mention of Victoria, Australia — as such it represented a probable real-world dataset and was a performance challenge for the classifiers. There were 90 VAEM and 524 non-VAEM in the imbalanced test dataset. Using this dataset for testing across all experiments allowed a fair comparison between the models when they were re-evaluated with more data and compared with their earlier performance. For the second phase of classification, a larger, balanced, test dataset of 828 documents was randomly extracted from the new records that were introduced in the second phase dataset, being 0.05 of the data. It was comprised of 431 VAEM and 397 non-VAEM, a 52.1% ratio, which left 9,564 VAEM and 9,685 non-VAEM in the main dataset for training and validation. The ratio of VAEM to non-VAEM in the main dataset was 49.7%. The Phase Two datasets are listed in Table 9.

Table 9: Phase Two datasets

Label	Main dataset		Large test data		Imbalanced test data	
Label 0 - VAEM	9,564	49.7%	431	52.1%	90	14.7%
Label 1 - Non VAEM	9,685		397		524	
Dataset Totals	**19,249**		**828**		**614**	

During training, the datasets (apart from test data) were first shuffled, then split into training and validation data with an 75/25 ratio.

4.8 Chapter 4 summary

This chapter first described the Twitter data that was collected over almost the entire 2018 year, using a search pattern with the Twitter API. This included an examination of examples of the vaccine-related tweets that were gathered. The chapter then discussed standard pre-processing techniques that are used to prepare text for machine learning, which included tokenization, removing unwanted tokens, and adding features. The rest of the chapter described the datasets that were assembled over two data collection phases, and how these were used for topic modelling and classification.

5 Topic modelling

5.1 Chapter overview

The previous chapter described how the text was prepared for topic modelling. This chapter discusses how topic modelling was used in this research. Topic modelling is a machine learning technique used to reveal distinct themes that each convey similar semantic meaning in a corpus of documents (Section 2.6.1), rendering these themes as different topics. Topic modelling can be used as a first step for exploring textual data (Paul & Dredze, 2017). Manual inspection and annotation of large corpora is very difficult, but topic modelling is used to automatically detect words that can help to identify a corpus's topics and which documents contain those topics.

Evaluation of topic models is based both on manual inspection of the key words of the model's topics and on using automated intrinsic and extrinsic measures - as explained in the "topic modelling evaluation measures" subsection of Section 2.6.1. The most useful evaluation measures are those that are aligned with the task that topic modelling is being used for. The primary goal of this research was to identify effective techniques for isolating VAEM to enable their extraction from the rest of the texts, so topic modelling was applied with a customized scoring technique to help identify which topics contain the most Vaccine Adverse Event Mentions (VAEM). This approach enabled an automated assessment of the optimum number of topics in a model that would result in one topic that best concentrated VAEM, then that topic was used to isolate or filter VAEM from the rest of the vaccine-related tweets.

Topic modelling also assisted the secondary objective of the research - to identify topics that could be used to form a taxonomy of vaccine-related Twitter posts.

5.1.1 Topic modelling algorithms

Latent Dirichlet Allocation (LDA) based models are generative probabilistic models and are presently considered a state-of-the-art method for topic derivation (Nugroho et al., 2020). They work on the assumption that each document can be represented by distribution over topics and each topic by distribution over words.

Dirichlet Multinomial Mixture (DMM) based models assume that each text can be described by one topic; DMM is applicable for modelling short texts and is an increasingly popular approach for topic modelling on Twitter data. For example, Surian et al. (2016) used DMM models to categorize opinions about HPV vaccine on Twitter data.

In the context of the main research objective to detect vaccine adverse event mentions in social media, several topic modelling algorithms were evaluated to determine a pragmatic approach that revealed a viable subset of VAEM-containing tweets from a corpus of vaccine-related tweets (Khademi & Haghighi, 2019). The following topic models were evaluated:

- LDA based model Gensim (Řehůřek & Sojka, 2010) developed in python programming language using online variational inference (C. Wang et al., 2011) and online learning (Hoffman et al., 2010).
- LDA based model MALLET ("MAchine Learning for LanguagE Toolkit") (Mccallum, 2002) via a Gensim wrapper using collapsed Gibbs sampling method (*Latent Dirichlet Allocation via Mallet*, 2020).
- DMM model from the jLDADMM library (hereafter just referred to as DMM) developed by Nguyen (2018). An enhanced version of the DMM model called Latent Feature DMM (LF-DMM) (Nguyen et al., 2015) includes the assessment of word vectors, this was also evaluated but found to be of no benefit for this dataset.

After inspecting the Twitter data it was evident that a tweet generally contained only one significant topic. Therefore, to evaluate the effectiveness of LDA-based models they were treated as if assigning only one topic per document by considering their dominant topic only, effectively producing the same output as the DMM models. This enabled straightforward comparisons between the topic models.

Coherence is a scoring method to measure the internal consistency of words allocated to a topic, and is normally used as a guide to decide on an optimum number of topics (Newman et al., 2010). CV coherence-scoring (Röder et al., 2015) was determined to be the most useful coherence scoring approach because of its understandability and was available with the Gensim LDA and MALLET models. Coherence scoring of the DMM model was not included in the model. It required a gold-standard set of labelled data, where a ground-truth label file must contain the "golden label" (i.e. verified by a system expert) of every document in the corpus (Nguyen, 2018). Since this would require labelling the entire corpus, coherence scoring was not able to be used with the DMM models.

5.1.2 Topic modelling data

The topic models were trained on data assembled during the first data collection phase, and later evaluated on data made available in a second data collection phase. Topic modelling data preparation and datasets are described in Section 4.4.

5.1.3 Labelling for topic model scoring

To be able to understand how topic modelling distributed tweets into topics, it was decided to manually pre-label some tweets, to be able to observe how the topic models categorized them and to understand which topics identified the VAEM. Random selections of tweets were inspected, and each tweet was assigned a label, based on the author's understanding of the major category the tweet belonged to. This process continued until no more major categories were identified and the categories that were most like VAEM were well represented. The result was that 1,400 tweets were labelled with 10 labels, with 222 of the labels assigned to the VAEM topic. These are listed in Table 10.

The domain expert also examined hundreds of examples of the tweets and verified that the annotated tweets were appropriately categorized. His observations and guidance at this stage of the study formed the ground-truth for what constituted VAEM and was used as the guideline for later annotation. All other tweets were labelled with a default value of 99. A scoring approach (see Section 5.2) used the labels to track when VAEM-labelled posts were identified in the models' topics.

Table 10: Manually assigned topic labels

Label	Topic
0	Vaccine Adverse Event Mentions (VAEM)
1	Enquiries / Discussions mentioning vaccines
2	Obvious sentiment against vaccines — anti-vax
3	Sentiment against anti-vax viewpoints, pro vaccines
4	Statements from vaccine related organizations
5	News articles and other factual or fake news
6	Nonsense / Spoof hijacking Vaccine meme
7	Everything else
11	Animal related
12	Advertising
99	Un-labelled data

Label 0 was given to the posts having vaccine adverse event mentions, which was the main topic of interest for the research. The zero value was used for the VAEM because it was the first number in a Python sequence, and this convention was followed also in subsequent labelling for classification. Generally, VAEM are effects like sore arms that and being currently

experienced, but discussions about similar past events were not excluded. Many of the examples are in the context of receiving a flu shot. Table 11 contains three examples of VAEM and three of tweets that are not VAEM.

Table 11: VAEM tweets examples

Vaccine adverse event mentions (label 0)
Vaccinations suck when they make ur baby sick :(
I got a flu shot Tuesday and my arm seriously hurts so bad.
I got my flu shot on March 1 (03/01). By Friday (03/09), I had noticeably developed a runny nose, cough and sore throat.
Not adverse event mentions (label 1)
i have to get my flu shot today but i hate shots, ik it's not a big deal but they freak me out, same with blood tests ugh
Gotten the flu shot every year, never had the flu. Quit blaming it on the vaccine people.
I'm currently dying with the flu but I refuse to get a flu shot cause shots make you sicker so doctors and pharmaceutical companies make more money

Note that although many of the adverse event mentions seem poorly worded and trivial and are expected side-effects such as a sore arm, these had been confirmed by the domain expert as vaccine adverse event mentions and the kind of data this study should try and obtain. That is, the study should find low-level personal narratives around experiences of adverse health events regarding vaccines, as this is the data that is missing from the formal reporting systems. This data can be used to measure trends, despite its seemingly inconsequential nature. The non-VAEM tweets are similarly worded but do not have semantically clear evidence of a personally experienced reaction to a vaccination. The challenge of the research is to discover and describe techniques that are able to discern the difference between these two very similar types of texts, and to accurately deliver a high volume of VAEM.

5.2 Topic modelling scoring method

Each of the topic modelling methods were evaluated against a range of topics, incrementing by 1, utilizing a loop with automated scoring. For each iteration, CV coherence was calculated where available (with the Gensim and MALLET models), but the main scoring technique used a form of F-Scoring that showed how many of the pre-labelled VAEM tweets were being grouped into a single topic. The main objective was to find the number of topics that resulted in the best ratio of VAEM to non-VAEM in one or two topics. At the same time, there was a preference for the smallest number of topics that still satisfied the main objective, as that was better for overall topic understanding. Experimentation showed that the ability for topics to

concentrate VAEM degraded after around 20 topics, and that they began to clearly differentiate VAEM from around 4 topics. Therefore, a range of 2 to 22 topics was used to allow a discernment of trends leading to and from the most significant points. For conciseness, some of these topic numbers are not shown in the figures below.

5.2.1 Calculating F-Scores

The scoring method used was standard F-Scoring (Manning et al., 1999) on the single best topic that captured the most VAEM (recall in F-Score terminology) with some degree of precision. Since identifying the greatest number of VAEM-containing documents was more important than unambiguously isolating the VAEM-containing documents, in standard F-Scoring terms recall was more important than precision. Therefore, F1-Beta scoring was additionally used, and named the Adjusted F-Score. F1-Beta scoring allows an adjustment of the importance of recall and precision. Experimentation revealed that a Beta of 1.3 was optimal — it increased the importance of recall but still accounted for precision, anything lower was not too different from standard F-Score, and higher values did not give enough importance to precision. Table 12 summarises the scoring metrics.

Table 12: Precision, Recall, F-Score and Adjusted F-Score

Precision	Proportion of VAEM captured in a topic vs total records in the topic
Recall	Proportion of VAEM captured in a topic vs the total number of VAEM
F-Score	2 x (Precision x Recall) / (Precision + Recall)
Beta β	1.3
Adjusted F-Score	$(1 + \beta^2)$ x (Precision x Recall) / ($(\beta^2$ x Precision) + Recall)

Precision and recall scores were based on how many of the labelled VAEM documents were identified in the single best topic, which naturally varies as topic counts are adjusted. That is, documents manually labelled as containing a VAEM were counted per topic and compared with the total VAEM count and the counts of other labelled documents within the topic. The best performing topic was taken as the scoring topic, as that indicated when the model was finding the characteristics that made VAEM tweets unique and placing these into one topic. However, if the top topic was not performing above a recall threshold of 0.6 and combining the top two topics produced a better score, then the combination of the top two topics was used.

As earlier described, the labelled documents were a small subset of 1,400 records, 222 were labelled with VAEM (label 0), 367 were labelled as discussions (label 1) etc. F-Score and

Adjusted F-Score were calculated from these scores, and topics were ranked by the Adjusted F-Score.

For instance, consider the 9-topic model in Table 13, and how it locates 221 of the VAEM documents into one best topic, which is Topic 8. At the same time, it also locates other labelled documents into Topic 8, resulting in a total 505 labelled documents. The precision, recall, and F-Scores attributed to Topic 8 are calculated from the relationships between the 221 VAEM and the total of 222 VAEM, and between the 221 VAEM and the total of 505 labelled documents in the topic.

Table 13: 9-topic DMM model

	Labels										
Topics	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**11**	**12**	**Total**
08	221	180	13	15	3	0	72	0	1	0	**505**
07	1	46	35	131	0	0	234	0	1	0	**448**
01	0	16	1	1	14	5	4	1	0	1	**43**
02	0	21	0	1	6	1	24	3	2	1	**59**
03	0	31	61	15	0	8	46	2	0	0	**163**
04	0	25	3	4	3	4	8	0	0	1	**48**
05	0	13	0	1	0	3	5	0	0	1	**23**
06	0	17	1	2	2	2	4	0	0	0	**28**
09	0	18	21	26	1	9	8	0	0	0	**83**
Total	**222**	**367**	**135**	**196**	**29**	**32**	**405**	**6**	**4**	**4**	**1,400**

To aid understanding, Table 14 summarizes the counts in the 9-topic model. For precision, the VAEM count in the Best Topic (Topic 8) is divided by the count of all labelled records in the topic, which is 221/505, a score of 0.438. For recall, the VAEM count is divided by the total VAEM count, which is 221/222, a score of 0.995. The F-Score and Adjusted F-Score calculations are calculated from these.

Table 14: Scoring of the 9-topic DMM model

	Labels			
Topics	**VAEM**	**All Other Labels**	**Total**	**Precision**
Best Topic	***221***	265	***505***	**= 220/485**
Other Topics	1	913	914	**0.438**
Total	***222***	1178	1419	
Recall	**= 221/222**	**F-Score**	**Beta**	**Adj F-Score**
	0.995	**0.608**	1.300	**0.675**

The 9-topic model identifies 99.5% of VAEM-labelled documents into one topic, i.e., a recall of 0.995. However, it also brings in other labels, so the precision is somewhat low at 0.438, and consequently the F-Score is 0.608. Adjusted F-Score using a beta of 1.3 favours recall and so the Adjusted F-Score is higher at 0.675 to indicate the importance of recall.

The equivalent figures from a DMM 14-topic model are shown in Table 15. Although the recall is marginally lower due to 1 labelled VAEM being moved into another topic, the number of non-VAEM in the best topic has been reduced, resulting in greater precision and consequently better F Scores — so it is the preferred model.

Table 15: Scoring of the 14-topic DMM model

	Labels			
Topics	**VAEM**	**All Other Labels**	**Total**	**Precision**
Best Topic	***220***	265	***485***	**= 220/485**
Other Topics	2	913	915	**0.454**
Total	***222***	1178	1400	
Recall	**= 220/222**	**F-Score**	**Beta**	**Adj F-Score**
	0.991	**0.622**	1.300	**0.688**

Figure 7 charts the changes of recall, precision, F-Score and Adjusted F-Score as the topic count increases and helps to illustrate the most suitable model as being that with the best combination of recall, precision and resulting Adjusted F-Score. Topic count also plays a part when there is a tie in the scores and a smaller topic count yields more interpretable topics. When examining the chart, the reader should observe the relationship between the Adjusted F-Score and F-Score lines, and how recall fluctuates in relation to precision.

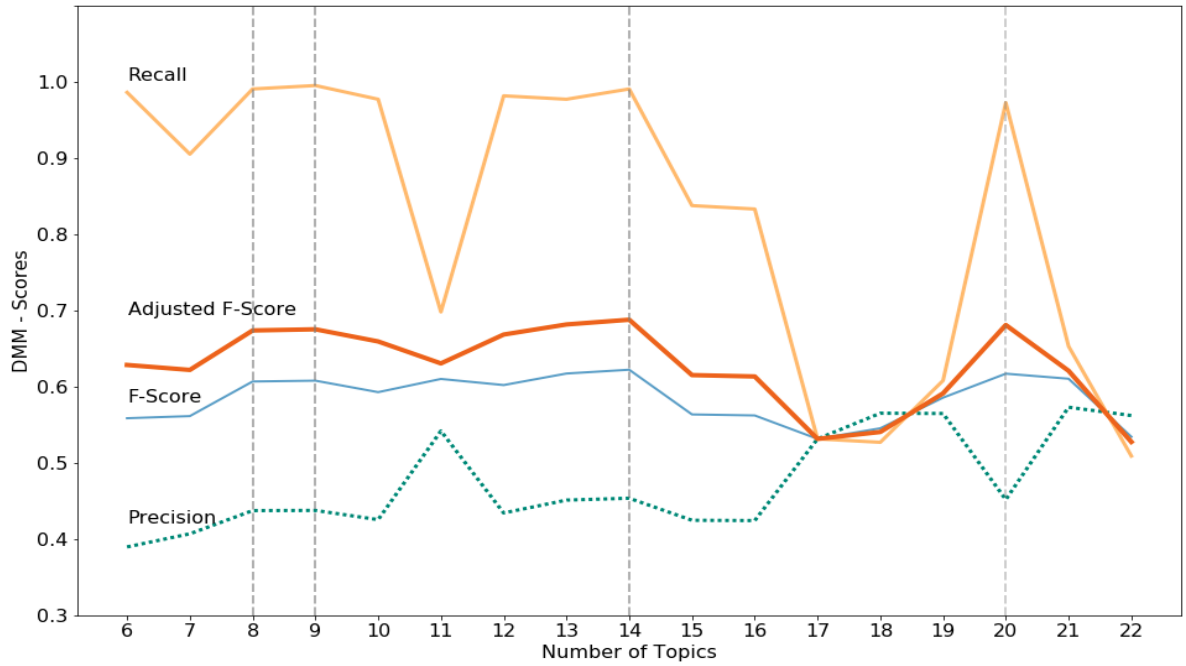


Figure 7: Scoring of DMM model per topic count

The chart shows that recall and precision fluctuate in an inverse relationship. This is due to the measurement of these being taken just from the best topic, and as the models are trained with increasing topic numbers then at times the VAEM are split over two or more topics, which leads to poorer recall at that point, and usually accompanied by greater precision. When VAEM are gathered into mostly one topic then recall improves, but at the expense of precision. The ideal situation is one with a very high recall and the best possible precision for that recall. Adjusted F-Score is used to favour the recall, and to allow for arbitration between recall and precision.

For instance, consider Topic 11 on the chart. At that point, the F-Score is the highest it has been to that point, due to a very high precision, but this is just because the model has split the VAEM, and the recall is correspondingly low. Although F-Score has increased, adjusted F-Score has markedly decreased, and since recall is more important than precision, should be used to decide about that model. F-Score and Adjusted F-Score are more in harmony when both precision and recall are aligned but Adjusted F-Score allows a clearer picture of the best performing topic models that favour recall. As shown in Table 15 and in the chart, adjusted F-Score is marginally better at 14 topics, with the 8, 9, and 20 topic models also contenders for the best Adjusted F-Score. Comparing the top 3 models, the 9-topic model has slightly better recall but less precision, the 20-topic model has the second-best precision but the least recall of the three models, and the 14-topic model has the best combination of recall and precision — both of its F-Score and Adjusted F-Score are better than the other models' scores.

Considering the 20-topic model is useful to clarify why it was preferable to find the model that minimized the splitting of VAEM while achieving a high as possible precision. The best topic in the 20-topic model contained 216 labelled VAEM, with 262 other labelled posts in the topic — its precision was therefore 0.452, better than the 9-topic model (at 0.438) and slightly less than the 14-topic model (at 0.454). However, its recall of 0.973 was lowest of the three models, and it had its 6 remaining VAEM spread over 5 other topics. This was an additional 4 VAEM outside of the best topic when compared to the 14-topic model, and 5 compared to the 9-topic model.

Table 16 illustrates these points using the numbers from 20-topic model. The rows of the table show the VAEM vs other posts per topic, with accumulating totals, and with precision, recall, and F-scores corresponding to these. The best topic's F-Score of 0.617 and Adjusted F-Score of 0.681 outperform those of the 9-topic model show in Table 14, which were 0.608 and 0.675. But these better scores come with a lower recall, as only 216 VAEM are captured by the topic, instead of 221 in the 9-topic model, and the gain in precision does not compensate adequately for the loss of recall. Adding just 2 more VAEM to get closer in recall would require combining the first two rows, then the precision plummets to 0.265, the F-Score is reduced to 0.417, and the Adjusted F-Score to 0.489. Clearly it is better to find a model count that balances precision and recall, to obtain as many VAEM as possible in one topic, with some precision.

Table 16: Relationship of scores to split VAEM in the 20-topic DMM model

				Accumulating Totals					
Topic	**VAEM**	**Other**	**Total**	**VAEM**	**Total**	**Precision**	**Recall**	**F-Score**	**Adj F-Score**
06	216	262	478	216	478	0.452	0.973	0.617	0.681
03	2	344	346	218	824	0.265	0.982	0.417	0.489
01	1	145	146	219	970	0.226	0.986	0.367	0.438
18	1	30	31	220	1001	0.220	0.991	0.360	0.430
09	1	14	15	221	1016	0.218	0.995	0.357	0.427
12	1	59	60	222	1076	0.206	1.000	0.342	0.412
Other	0	324	324	222	1400				
Total	222	1178	1400						

The spread of VAEM over other topics could only be considered useful if it were required to discern some differences in the VAEM that the topic model had identified. However, the goal was to be able to extract the best balance of VAEM to non-VAEM, with no distinction required in discerning the type of VAEM. Given that both the 9-topic and 14-topic models did better at gathering VAEM into one topic, they were preferred. Even if the best, 14-topic model, had not been available, it would still have been better to pick the 9-topic model over the 20-

topic model, despite the 20-topic model outscoring the 9-topic model, and deal with the poorer precision afterwards. Unless of course, it had been preferred to gain precision at the expense of recall at this point, *which was not the case*.

In conclusion, the overall low precision in all these models was not a reason for concern, as the objective in using the topic models was to identify topics that contain the most VAEM, ideally with fewer of other topics, so that a filtered set of data could be obtained for labelling and handling by the subsequent classification process. In other words, precision was only important because it helped identify models that had a strong recall but with not too many other competing texts, which was when the topic model was acting as a most effective filter. Where models' scores were nearly equal, preference was given to recall, and then to a lower total number of topics, as a lower number of topics was more comprehendible.

5.2.2 Coherence

Coherence acts as an intrinsic measure of the human interpretability of topics (discussed in the Topic modelling evaluation measures section of the literature review), so where it was available (Gensim and MALLET) it was assessed to see if it could help with choosing the optimal number of topics. However, it was found that coherence was very often out of sync with the best scores using the F-Scoring approach. For instance, the difference between Adjusted F-Scores and traditional topic coherence obtained on the Gensim models is illustrated in Figure 8; for both measures a higher score is better.

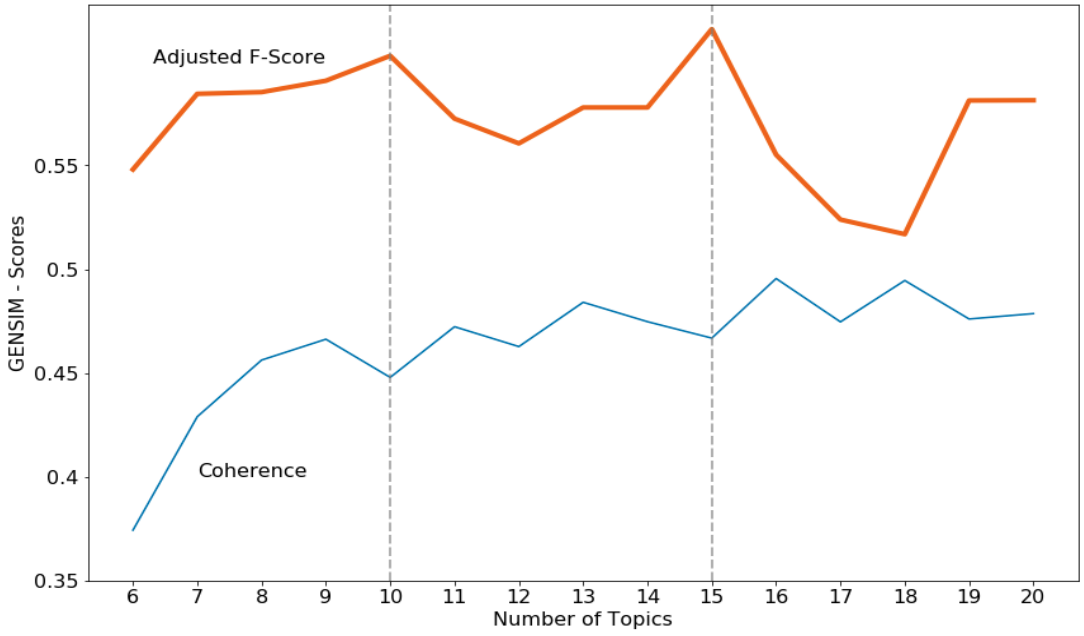


Figure 8: Coherence vs F-Score per topic count

The points where Adjusted F-Scores are highest are where the model is identifying the best numbers of topics to isolate VAEM-containing documents, but at these points, coherence has always decreased. As it had been decided that the best number of topics should be based on finding at least one topic that clearly identified most of the VAEM, with a preference for topics that were also human interpretable, coherence was only examined as a general trend.

5.3 First stage of topic modeling

In the previous section topic modelling was introduced, with an explanation of the scoring calculations used to determine the best topics for isolating vaccine adverse event mentions. This section compares the results of applying this scoring approach over a range of topic numbers using Gensim LDA, MALLET, and DMM topic models. Figure 9 compares the adjusted F-Score for the three models over a range of 4 to 22 topics. As described previously, the scores were calculated as the Adjusted F-Score of the single best topic at each iteration — to help determine which topic model might most usefully identify potential vaccine adverse event mentions into one topic.

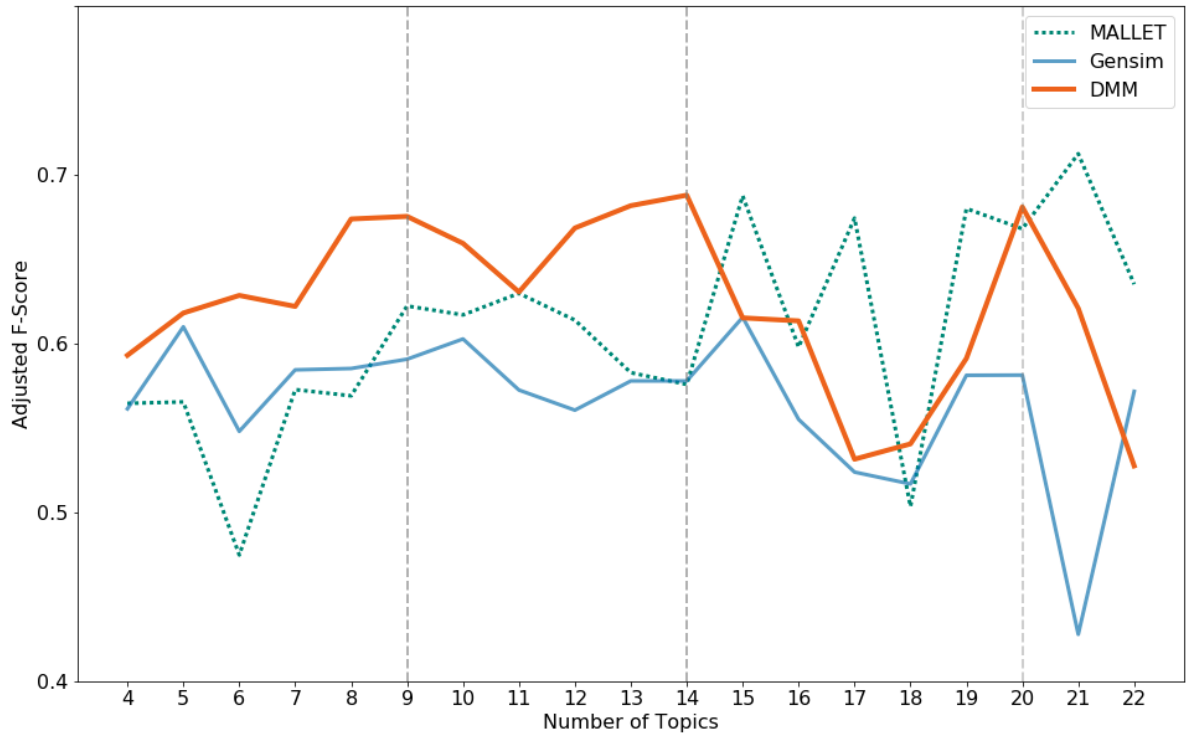


Figure 9: Comparison of Adjusted F-Score vs topic counts

All the models showed peaks in Adjusted F-Score at around 9 or 10 topics, then at 14 or 15 topics, and again at 20 or 21 topics. The best model for identifying VAEM up until 14 topics was the DMM model and it was clearly the preferable model based on its performance in this lower range of topics, where fewer topics are more understandable and therefore preferred. It

performed well also at the 20-topic mark, while none of the models behaved consistently well between 15 and 19 topics.

5.3.1 DMM model

The DMM model recall and precision are shown with Adjusted F-Score in Figure 10. It performed admirably at 8 and 9 topics with a very high recall and although the precision was on the low side, the combination of the adjusted F-Score and the understandability of the topics indicated that one of these was a candidate for selecting a subset of VAEM-containing posts. The 9-topic model had the most understandable separation of topics while also having a reasonably high Adjusted F-Score. Slightly higher Adjusted F-Scores were obtained at 14 and 20 topics in the DMM model - the 14-topic model performed best for the specific task of identifying a subset of potential vaccine adverse event mentions, but its topics were not as understandable as those of the 9-topic model.

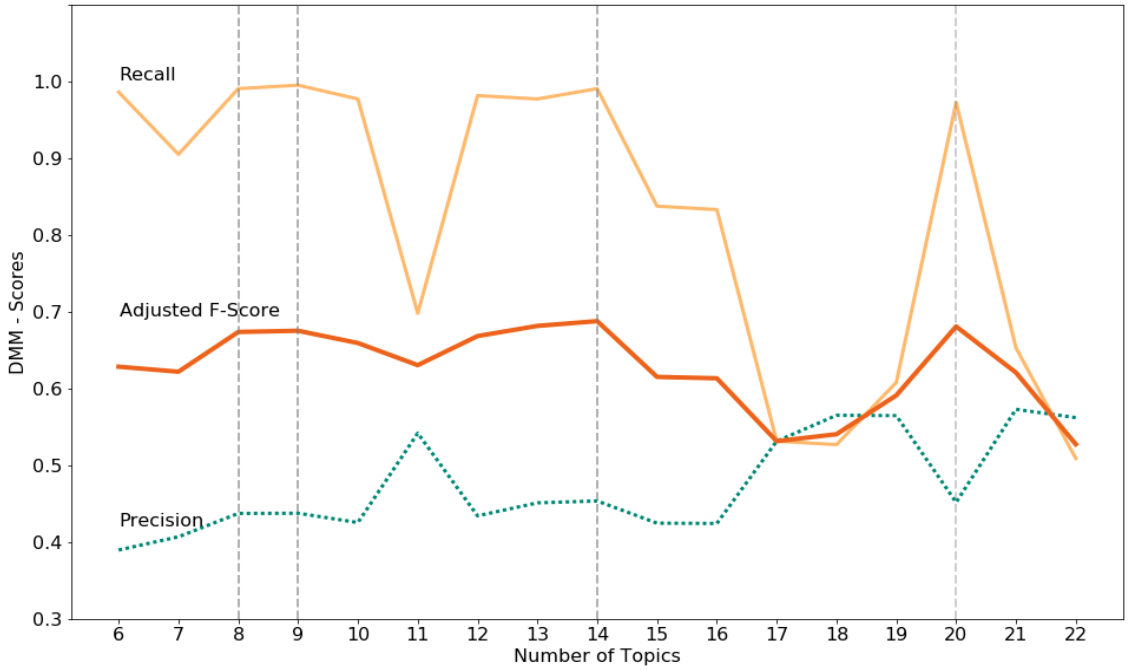


Figure 10: DMM model scores

5.3.2 MALLET model

The MALLET model (Figure 11) displayed increasing Adjusted F-Scores, but only competed with the DMM model from the 15 topics mark. However, from this point also its Adjusted F-Score exhibited a wildly fluctuating pattern. This reflects an observation made when using MALLET, that in the process of more-or-less evenly distributing documents over the allocated topic numbers, it would tend to distribute VAEM containing documents over many groups. At

the points where MALLET separated the target documents into many topics it would perform badly due to a lack of recall, at the times when it gathered the targets together into fewer topics its performance was let down by a lack of precision. Besides this problem, the key words MALLET identified per topic were not as understandable as those delivered by other models — these differences are explored in Appendix A.

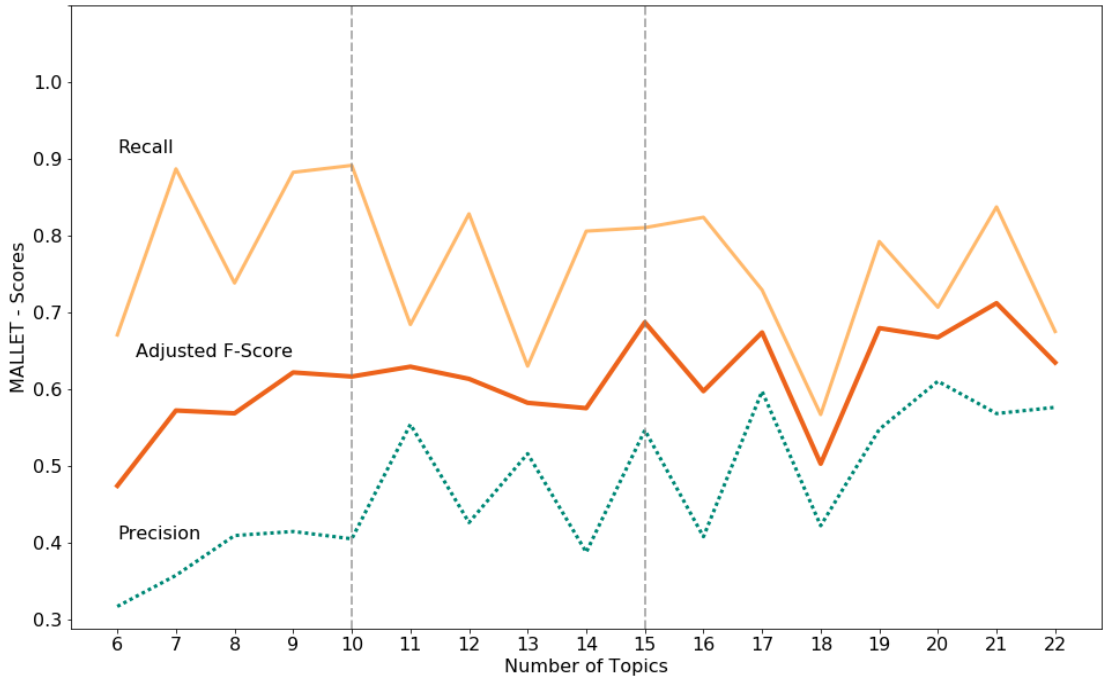


Figure 11: MALLET model scores

5.3.3 Gensim model

The Genism model depicted in Figure 12 was a steadier performer. Starting with a reasonable precision and recall and at 10 topics the topics were very understandable and preferred to those obtained with the Gibbs Sampling used by MALLET, while having a similar performance at that point.

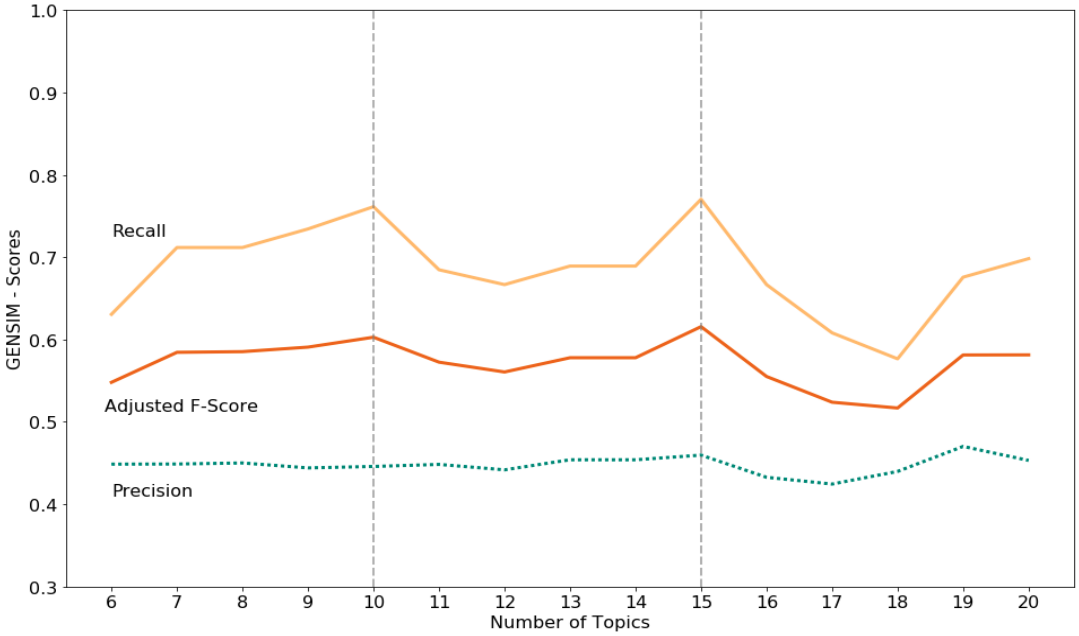


Figure 12: Gensim LDA model scores

5.3.4 Summary of the best scoring topic models

Table 17 summarizes the best scoring topic models, based on Adjusted F-Score on the best topic but also with consideration to those that identify the most VAEM (i.e. recall). Precision is the ratio of *Label 0* to *Total (documents) in topic*. Recall is measured as the ratio of *Label 0* to the *Total of Label 0* in the models, which was 222.

Table 17: The best scoring topic models

Topic Model	Label 0	Total in topic(s)	Precision	Recall	F-Score	Adjusted F-Score
Mallet 9	196	472	0.4153	0.8829	0.5648	0.6223
Mallet 10	198	488	0.4057	**0.8919**	0.5577	0.6170
Mallet 15	180	329	**0.5471**	0.8108	**0.6534**	0.6876
Gensim 10	169	379	0.4459	0.7613	0.5624	0.6028
Gensim 15	171	372	0.4597	0.7703	0.5758	0.6156
Gensim 18 *	206	530	0.3887	**0.9279**	0.5479	0.6122
DMM 9	221	505	0.4376	**0.9955**	0.6080	0.6754
DMM 14	220	485	0.4536	0.9910	0.6223	**0.6880**
Total Label 0	**222**		* Combined		Beta	1.3

The best model identified by Adjusted F-Score is the DMM 14-topic model with a score of 0.688, which achieves this because of its precision of 0.454 combined with its high recall of 0.991. This is in preference to the DMM 9-topic model - which has the highest recall of 0.996, but with less precision. Although nine topics was more understandable than fourteen and the high recall is significant, for filtering VAEM the almost identical recall with greater precision of the 14-topic model was preferred. The actual difference in counts of VAEM between the two DMM models was *one*.

If standard F-Score were to be used then the best topic model would be the MALLET 15-topic model with an F-Score of 0.653, but it achieves this because of its much higher precision of 0.547. However, its recall is relatively low at 0.811 compared to the other MALLET models, the Gensim LDA 18-topic model (when combined), and the DMM models.

Examined in more detail in Table 18, scoring of the best topic (which was Topic 13) of the DMM 14-topic model shows that it captured 220 of the 222 labelled VAEM, a recall of 0.991. Although it was not very precise at 0.454 it was better than most other models in that regard, and so outperformed all other models because of its precision / recall combination.

Table 18: Counts and Scoring of the 14-topic DMM model

	Labels					
Topics	**VAEM**	**Other Labels**	**Total**	**Precision**	Unlabelled	**Total**
Best Topic	**220**	265	**485**	**= 220 / 485**	37,117	**37,602**
Other Topics	2	913	915	0.454	290,305	291,220
Total	**222**	1,178	1,400		327,422	328,822
Recall	**= 220 / 222**	**F-Score**	**Beta**	**Adj F-Score**		
	0.991	0.622	1.300	0.688		

When consideration is given to recall over precision (where precision over recall was a major contributor to the Adjusted F-Score), the best models were still the DMM 14-topic model as its recall was only slightly less than the DMM 9 topic model; the Gensim combined 18-topic model; and the MALLET 10-topic model. Data from these topics was extracted for a second stage of topic modelling, which was expected to provide greater precision. This is explained in Section 5.5.

5.3.5 First Stage Topics keywords

The 14-topic and 9-topic DMM models were very similar, with top keywords of *"get, flu, shot, not, do, be, go, vaccination, year, sick, take, never, shoot, today, people, time, day, vaccinate, feel, need"* in the VAEM topic of the 14-topic DMM model. This was the preferred model as it concentrated the most VAEM into one topic, while splitting off non-VAEM into other topics.

The topic keywords of the 10-topic Gensim model were arguably the clearest set of different topics, and all the topics were understandable and well differentiated. The top 20 words of the VAEM topic in the Gensim model were *"flu, get, shot, year, not, shoot, sick, still, time, last, season, never, today, be, go, take, bad, arm, week, feel"*. The models that did not perform well tended to split VAEM into several topics and furthermore produced spurious topics that could not be meaningfully categorized. These are discussed in detail in Appendix A.

5.4 Taxonomy

A taxonomy is defined as a form of classification for structuring and organizing entities (Nickerson et al., 2013). Table 19 shows the taxonomy that was developed to describe the main topics in the tweets.

Table 19: Taxonomy of vaccine related Twitter posts

Subject	Description
Vaccine Adverse Event Mention (VAEM)	Personal mentions of experiencing an adverse event after receiving a vaccine
Personal Health Mention	Personal mentions of experiencing health issues but not VAEM
Discussions	Enquiries / Discussions / Complaints mentioning vaccines; can be emotional, sensational or neutral, but not overtly pro or anti-vaccination
Pro-Vaccination	Sentiment or language against anti-vax viewpoints, pro vaccines, including promoting and advertising vaccines, can be implicit
Anti-Vaccination	Obvious sentiment against vaccines; anti-vax
Autism	All autism related discussions
HPV & Cancer	HPV and cancer-related vaccine discussions
Pets and Veterinary	Pet and animal related discussions, including what might be classed as VAEM had they related to human subjects
Trends and Outbreaks	Statements and headlines mentioning trends and outbreaks
Research and Studies	Mentions of new studies and research, science of vaccine development, including headlines mentioning research
News	News articles, headlines, and announcements. Statements from vaccine-related organizations
"The Vaccines"	Mentions of the indie rock group The Vaccines; these could be filtered out at data collection time

The taxonomy accounted for the topics found by the topic models, and so is more aligned to the patterns in the data compared to the manually determined label scheme that had been used for the topic model scoring. The taxonomy was derived by evaluating the most understandable topic models from the three topic modelling algorithms and by examining the data. As described in Section 5.3, in the course of topic modelling it was found that all the models produced understandable divisions around 9 or 10 topics, and that the resulting topics were somewhat consistent across the models (see Appendix A), so a combination of the suggested topics was used in the taxonomy.

The taxonomy accounts for how topic models see the data and contains distinctions that were not apparent when the data was initially examined and manually pre-labelled. For instance, HPV and cancer-related discussions were distinct topics in the models. Nevertheless, there is a reasonable alignment with many of the preliminary observations that were the basis for the initial 1,400 labelled records used for topic model scoring, especially around the divisions of VAEM-containing posts compared with discussions, and with the pro and anti-vaccination groups. The distinction between vaccine adverse event mentions and other personal health mentions was created manually to assist with analysis, the topic models did not see these as separate. Likewise, the last topic of the rock group "The Vaccines" was manually added, as it was easy to ascertain these posts, so it made sense to delineate and target them for removal — perhaps as part of the initial data processing, and certainly before classification. "Pets and Veterinary" was adopted as a label for the many posts that mainly discuss pet vaccinations, and was a topic found by the models. The taxonomy places all animal-related posts into this group, even if the text of the post is describing what would otherwise be considered a vaccine adverse event mention, as the research needed to focus on human subjects — but this merging of all animal-related posts was not present in the topic models.

Apart from the manually inserted or altered distinctions that were made to help with clarifying VAEM, the taxonomy does not differentiate topics any more than the topic models did. For instance, although the initial manual scheme identified "spoof" or fake posts as a separate area, the topic models did not find this distinction at around 10 topics, nor did it discern a separate "fake news" topic — so the taxonomy does not show these. The distinctive features of these kinds of posts might have been discernible if the topic numbers were sufficiently increased, but at the cost of decreased comprehensibility.

Since the aim was to choose the topic model design and number of topics that concentrated most labelled vaccine adverse event mention posts into one or two topics for a *binary*

classification task, this taxonomy was not applied to further labelling. However, it was used on some data samples to verify the effectiveness of the topic modelling approach — see Section 7.2.1. The Appendix B has a detailed view of the mapping between the topics' keywords and the topics of the taxonomy.

5.5 Second stage of topic modelling

The previous section described the process used to decide on the best topic models per topic model type, which identified a subset of VAEM containing tweets. This section deals with processing and evaluating the best three data subsets suggested by that analysis, again using the three topic models.

To this point, topic modelling had been used to capture nearly all VAEM posts into one topic, but relatively imprecisely — there were almost as many similar, but non-VAEM posts contained in the topic. After around fifteen topics any increase in precision was accompanied by a decrease in recall - the models would split the VAEM posts into different topics, and the models' topics were generally less understandable. Therefore, it was decided to do further topic modelling on just the extracted top topic (or combined topics) subset which already contained the VAEM posts.

To reiterate, to get to this stage Adjusted F-Score and recall were used to identify the best performing topic number (or numbers if combined), per model, per topic model architecture. The best topic(s) are hereafter named the "VAEM topic(s)". The best VAEM topics of the models were Topic 13 of the DMM 14-topic model, topics 3 and 6 combined of the Gensim 18-topic model and Topic 3 of the MALLET 10-topic model, and these were extracted for testing in the second stage, resulting in 3 data subsets.

Further topic modelling was performed on the three datasets, using each of the three model types, testing a range from 2 to 20 topics. Nine result sets were thereby obtained and compared with one another — the goal being to see how the 3 models performed on each of the 3 input datasets, and how their results compared with one another. The DMM model applied to the DMM data was the outstanding performer, resulting in an F1-Score of 0.821 and an Adjusted F-Score of 0.820 at 9 topics, due to a high precision of 0.829 and recall of 0.814.

The Adjusted F-Scores obtained in this second stage are considerably higher than the scores in stage one, 0.833 compared to 0.688 for the best scores. The results are demonstrated in Table 20, which shows the best results over the three datasets were all achieved by DMM, but the other two models' best scores were achieved over the Mallet dataset. A detailed analysis of the models' performances can be found in Appendix C.

Table 20: Second stage best scores per model & dataset

Topic Model	Dataset	Topics Count	Precision	Recall	F-Score	Adjusted F-Score
DMM	Mallet	18	0.778	0.869	0.821	0.833
DMM	DMM	9	0.829	0.814	0.821	0.820
DMM	Gensim	17	0.720	0.874	0.790	0.810
Mallet	Mallet	19	0.788	0.828	0.808	0.813
Gensim	Mallet	3	0.575	0.874	0.694	0.732

5.5.1 Second Stage topics keywords

Table 21 shows the first 20 keywords of each topic in the DMM 9 topic model. Many of the topics have the words “get”, “flu”, “shot” as their first three key words, which indicates how concentrated these topics are around the predominant cause of VAEM — getting a flu shot. All the topics contain varying amounts of VAEM, so contain many of the same words, but Topic 8 is the main VAEM containing topic, and its other key words “arm”, “today”, “feel”, “hurt”, together with “be”, “do” and “go”, indicate current activities around getting and reacting to a flu shot. The next two topics containing the most VAEM were topics 9 and 1. Topic 9 has many similarities to Topic 8, but “sick” is the only specifically vaccine reaction-related word in its top 20 keywords, whereas Topic 8 had “arm”, “hurt”, “sore”, and “pain”. The top 3 topics are the only ones containing the word “today” in their first 20 keywords, which corresponds with them being discussions of recent events.

Table 21: Second stage DMM 9 topic model keywords

1	get, not, vaccination, be, go, do, flu, shot, today, vaccinate, take, day, baby, need, time, doctor, give, say, know, feel
2	blood, not, baby, dog, give, be, get, would, cry, help, could, donate, sad, do, take, heart, sansa, someone, need, poor
3	flu, shot, get, not, take, shoot, do, eat, good, cold, day, make, sick, drink, vitamin, be, ginger, never, go, need
4	flu, shot, need, get, be, think, gravy, people, cold_bitch_think, not, cold, afraid, thot, know, change, college_graduation, cocaine, man, night, school
5	get, flu, not, shot, do, vaccinate, stay, sick, kid, home, hand, go, cough, take, people, know, healthy, catch, keep, be
6	flu, get, shot, not, do, people, be, go, say, shoot, sick, give, year, tell, take, need, know, die, never, work
7	flu, get, shot, not, year, never, sick, be, have, do, time, still, shoot, go, ever, feel, take, last, first, think
8	get, flu, shot, arm, not, vaccination, today, be, feel, hurt, go, sore, yesterday, still, do, day, needle, shoot, can, pain
9	flu, get, shot, not, year, be, day, go, do, shoot, sick, feel, week, take, time, good, still, today, work, have

5.6 Summary of the two stages of topic modelling

At the end of stage two, data was exported from the DMM 9-topic model. Topic 8, the *best topic* of the model, contained 81% of the VAEM and could be used for a final dataset, but data was taken from the top three topics, which were topics 8, 9, and 1 — together they captured 95% of the original VAEM (211 of 220). The rest of the pre-labelled VAEM was captured by Topic 7 (with 7 VAEM) and topics 3 and 4 (with 1 each).

Figure 13 summarizes the process of filtering the data through two stages of topic modelling, illustrating the filtering benefit obtained by focusing on the best topics from each stage. It shows that most of the relevant potential VAEM-containing documents have been identified and most of the irrelevant documents have been discarded. In the figure the large circles represent the entire data at each stage, the smaller circles within these represent the total labelled records, and within them are circles representing the VAEM-labelled records.

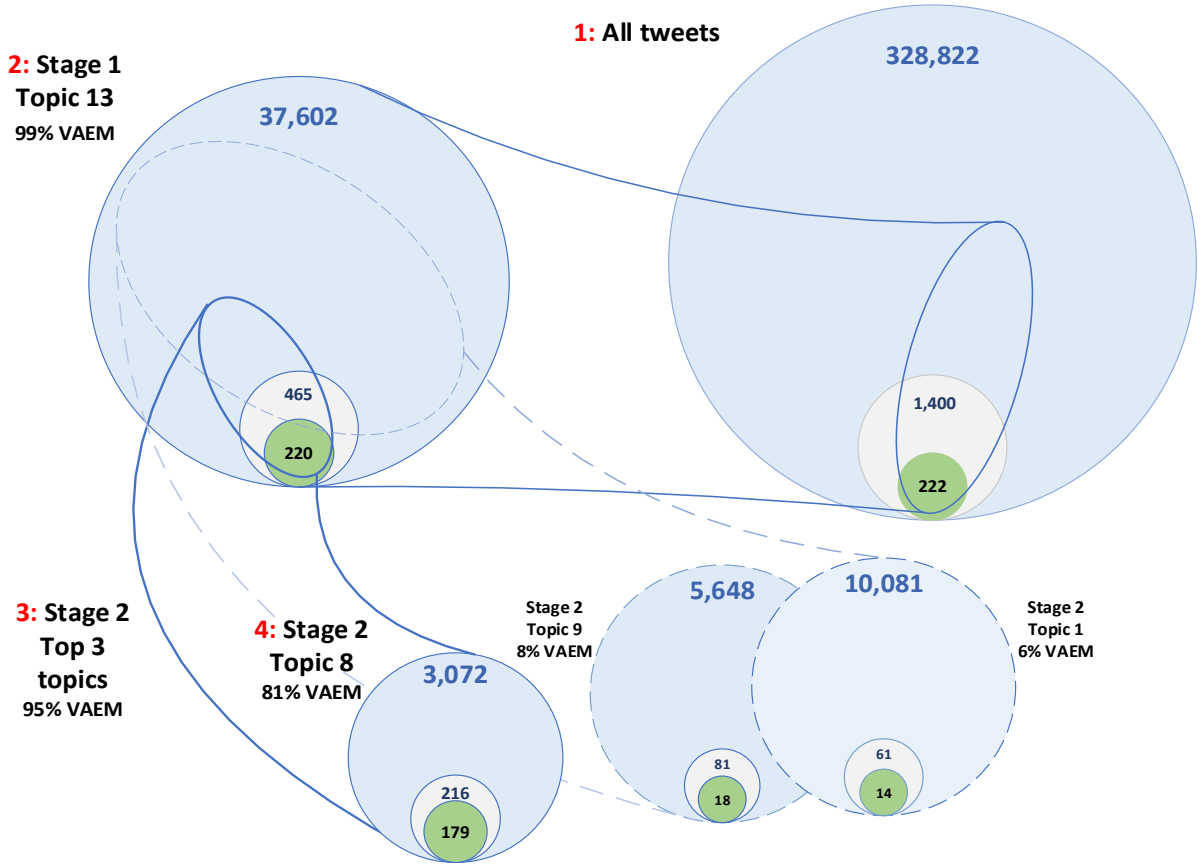


Figure 13: Two stages of topic modelling

The following four points are commentary for the numbered items in the figure:

1. Starting with all tweets: There were 328,822 documents. 1,400 were labelled, of which 222 were VAEM.
2. Stage one of topic modelling identified a single topic, Topic 13 of a 14-topic DMM model, which according to the labelled samples includes around 99% of the VAEM-containing documents (i.e. it included 220 of the 222 labelled VAEM). In doing so it has discarded 291,220 records, leaving only 37,602 — roughly 11% of the original records.
3. From this extracted stage-one dataset, a second stage of topic modelling was performed, resulting in a 9-topic DMM model which concentrated the VAEM texts further into a small number of topics. Note that the 9 topics provide an overlay for the stage-one records, taking a subset of them results in further filtering. The top 3 topics of this model were topics 8, 9 and 1 — when taken as a subset they filter out exactly half of the incoming records, including 4% of the VAEM. Consequently, 18,801 records would remain, which is 5.7% of the original records, containing 95% of the labelled VAEM (211 of the original 222 VAEM).
4. The best topic in stage two was Topic 8. It contained only 3,072 documents - around 1% of the original document count. According to the VAEM-labelled data this best topic contained around 81% of the VAEM (179 of 222).

To conclude, the evidence coming from the pre-labelled documents was that after a single stage of topic modelling 89% of the irrelevant tweets were discarded, to retain 99% of VAEM. Taking the top 3 topics from a second stage of topic modelling would dispense with 50% of the remaining records — discarding 94.3% of the original records to retain 95% of VAEM. Using just the best topic of the second stage topic model would result in a total effect of discarding 99% of the original texts to retain 81% of the VAEM. However, these were only *estimates* based on the small number of 1,400 pre-labelled tweets.

5.6.1 Verification of the best topic model

To verify that the best topics in the best scoring second stage model were indicative of its ability to identify the VAEM containing tweets, data from the three top topics of the 9-topic DMM model was manually labelled by the author, following the domain expert's guidelines. The uncertain cases were sent to the domain expert for clarification. As described above, these were topics 8, 9 and 1, and the criteria for choosing them was because together they captured 95% (211 / 220) of pre-labelled VAEM, while discarding a further 50% of the irrelevant tweets that were present in the data being processed by the stage two topic models.

The result of this labelling is shown in Figure 14. The top 3 bars in the figure show the proportions of VAEM (on the left) to non-VAEM (on the right) for each topic, and the percentage of total VAEM to the left of each bar. For instance, Topic 8 contains 1,300 VAEM and 1,772 non-VAEM, for a total of 3,072. Its VAEM is 75.1% of the 1,732 VAEM in all three topics. The bottom 3 bars display the topic's VAEM number and percentage of VAEM in detail on the left section of the bar, and on the right, they display the VAEM figure relative to the total tweets in the topic.

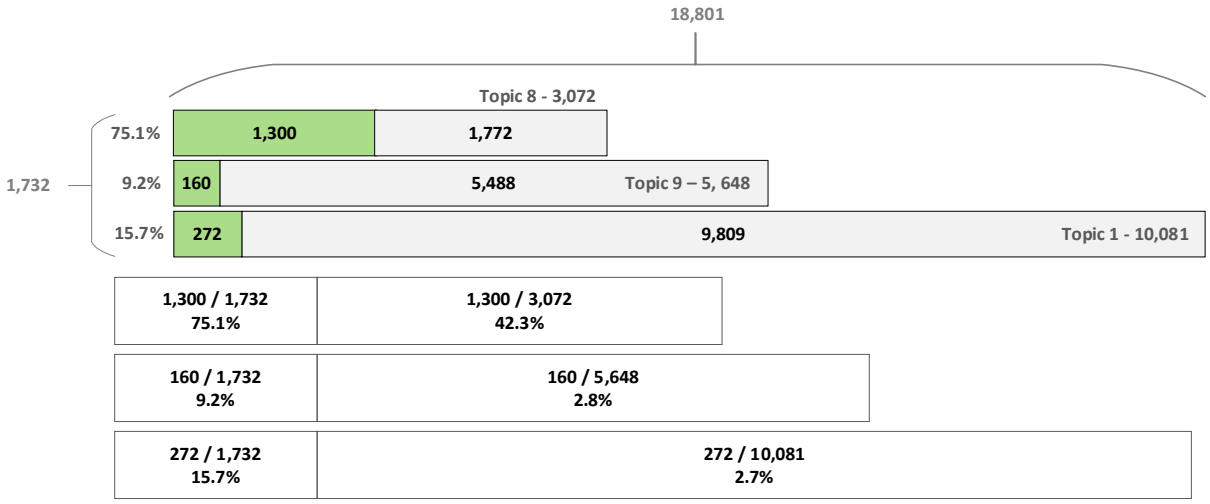


Figure 14: Labelled top 3 topics of second-stage topic model

After labelling, 1,732 VAEM-containing documents were identified in the top 3 topics. It was verified that Topic 8 contained the biggest percentage of VAEM, with 1,300 tweets, which is 75% of VAEM, rounded. The other top topics contained 9% and 16% (rounded) of the VAEM between them. These differed somewhat from the estimated 81%, 8% and 6% proportions which had been calculated from the small cohort of pre-labelled documents, but it supported the observation that Topic 8 was by far the best topic for capturing VAEM.

The proportion of VAEM to non-VAEM in each topic is significant, because the motivation behind topic modelling was to reduce non-VAEM so that a filtered subset of records could be obtained, ideally with a significant amount of VAEM in one topic, accompanied by as few non-VAEM as possible. Topic 8 clearly achieves this but given that it contained only 75% of the now fully labelled VAEM, it was justified to keep topics 9 and 1 to retain the remaining VAEM, and certainly for the initial requirement of training classifiers on the filtered data.

A prior observation was that Topic 7 might also contain significant VAEM, as it had contained 7 of the remaining 9 pre-labelled VAEM after counting topics 8, 9, and 1. To check this, it was also labelled, and was found to have only 94 VAEM tweets, which out of a total

topic count of 6,274 was around 1.5% of tweets in the topic. Since 94 was a small proportion of VAEM and adding over six thousand irrelevant records would have contributed additional noise to the data, it was decided to not include these records. Besides, expanding the dataset by including Topic 7 was contrary to the aim of effectively filtering the data via topic modelling.

5.6.2 "The Vaccines"

While labelling, 282 documents were found that were clearly discussing the indie rock group "The Vaccines", so these were assigned a new temporary label 88. On the basis that these were easily identifiable by the exact proper case expression "The Vaccines", they were eventually eliminated when creating the Classification datasets, reducing the total document count at that point to 18,519.

5.6.3 Final labels

Labelling of VAEM records in topics 8, 9, and 1 was completed. The eventual labelling aim was a binary label of vaccine adverse event mention or not. However, it was considered useful to retain some idea of what the non-VAEM documents contained, so some other labels were included. For Topic 8 the records were labelled as VAEM (label 0), discussions (label 1) or pet related (label 11), leaving only 30 unlabelled. Topics 9 and 1 were also fully labelled for any VAEM or pet related, but only around 800 more were labelled of the discussion-related documents. Anything else was left as unlabelled (label 99), but with the binary labelling scheme in mind these really counted as "other than VAEM" rather than their previous unknown status. The resulting 1,732 VAEM-labelled documents were confirmed by the domain expert. Table 22 show the final labelling over the 3 topics that were retained.

Table 22: Label distributions in top 3 topics

Label	Topic 8	Topic 9	Topic 1	Label Totals
Label 0 - VAEM	1,300	160	272	1,732
Label 1 - Discussions	1,700	65	754	2,519
Labels 2 to 6	6	7	73	86
Label 11 - Pets	13	28	800	841
~~*Label 88 - The Vaccines*~~	~~*23*~~	~~*58*~~	~~*201*~~	~~*282*~~
Label 99 - Unlabelled	30	5,330	7,981	13,341
Topic Totals	3,072	5,648	10,081	**18,801**
Totals excl The Vaccines	3,049	5,590	9,880	**18,519**

5.7 Evaluation

After labelling, the count of VAEM-containing tweets increased from 211 to 1,732. The ratios of VAEM to other data confirmed the effectiveness of the best topic (Topic 8) for identifying VAEM. Topic 8 contained 75.1% of the VAEM tweets identified in the three topics, and VAEM tweets were 42.3% of all the tweets in Topic 8. This is a lower proportion of VAEM in the topic than the 81.4% previously inferred using the partially labelled data.

The two other topics contained 15.7% and 9.2% of VAEM but the ratio of VAEM to other records in them was much lower at 2.7% and 2.8% respectively. It was evident that the two stages of topic modelling were very effective for obtaining a useful dataset of VAEM-containing tweets, which was amenable to further refinement via classification techniques. Table 23 summarizes the proportions of VAEM to other labels for the three topics, before and after labelling.

Table 23: Top 3 topics labelling summary

	Labels before labelling					After labelling				
Topics	**VAEM**	Other	Unlabelled	**Total**	% VAEM	**VAEM**	Other	**Total**	% VAEM	VAEM % of topic
08	179	37	2,856	3,072	81.4%	1,300	1,772	3,072	75.1%	42.3%
09	18	63	5,567	5,648	8.2%	160	5,488	5,648	9.2%	2.8%
01	14	47	10,020	10,081	6.4%	272	9,809	10,081	15.7%	2.7%
Sub-Total	211	147	18,443	18,801	95.9%	1,732	17,069	18,801		
Other topics	9	118	18,674	18,801	4.1%					
Grand Total	220	265	37,117	37,602						

Table 24 is a more summarized view of the proportions in the final labelled data with an emphasis on how well the main topic (Topic 8) does on concentrating VAEM, it contains 75.1% of VAEM, which are 42.3% of its total documents.

Table 24: Distribution of data in the VAEM topic

Topic	**VAEM**	**Other**	**Total**	
Topic 8	1,300	1,772	**3,072**	**42.3%**
Topics 9 & 1	432	15,297	**15,729**	
Total	**1,732**	**17,069**	**18,801**	
	75.1%			

The effectiveness of the two-stage topic modelling process was verified when the topic models were applied to another 400 thousand records after the second round of data collection,

and the exported data was labelled for classification. It was very clear that Topic 13 of the first stage identified virtually all the VAEM and that Topic 8 of the second stage continued to isolate around 75% of the VAEM. This is discussed in Section 7.2.2, “Verifying effectiveness with label distributions”.

5.8 Additional visualisation techniques

Apart from graphing the progress of F-Scores and coherence per topic, other visualisation techniques were used to assist with understanding the distinct and shared characteristics of topics, and if topics appeared to be sensible and explainable divisions of the discussions.

pyLDAvis

The pyLDAvis library (*pyLDAvis - Python library for interactive topic model visualization*, 2018) which is based on LDAvis (Sievert et al., 2014) was found to be very effective, as it enabled an interactive visual inspection of the topic and word distributions. For instance, Figure 15 shows a visualisation of the best stage two DMM model, with its main VAEM topic highlighted. Words like “arm”, “feel”, “sick” and “shot” appear more prominently in this topic than in any others, and if a word such as “arm” is selected then the chart shows its distribution in the various topics - Figure 16.

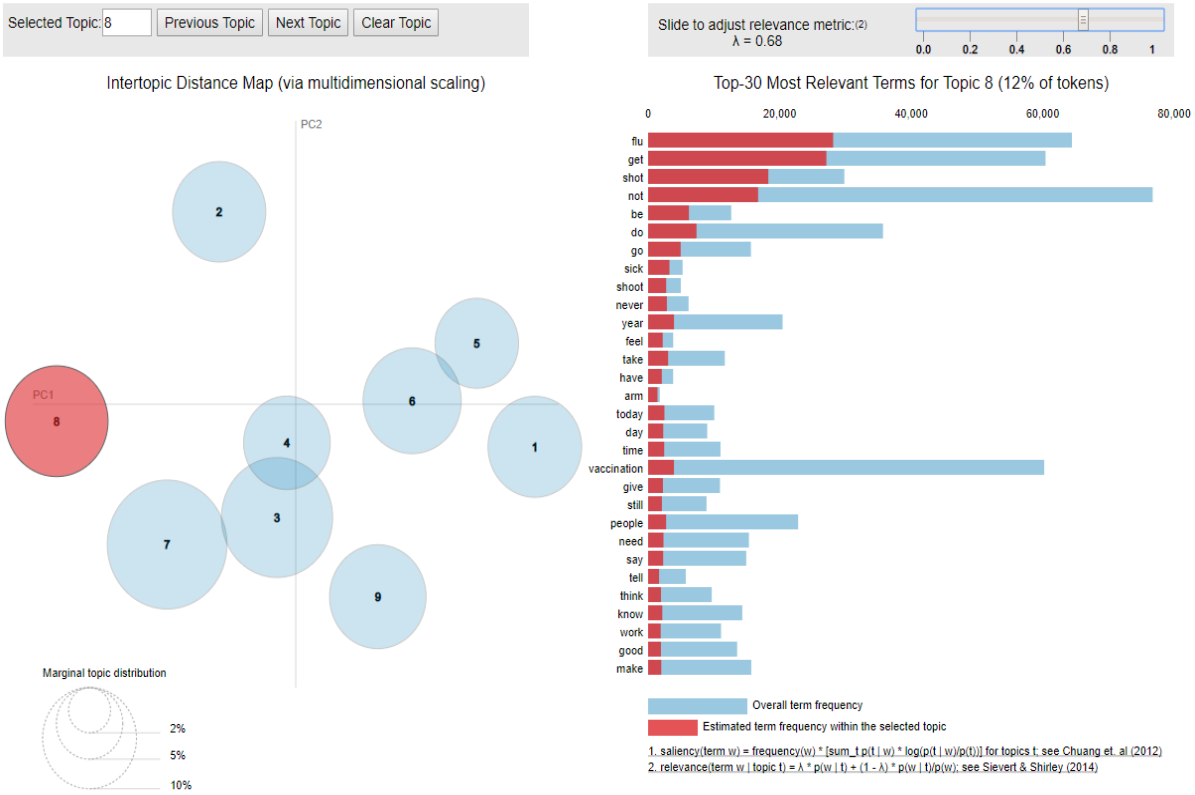


Figure 15: pyLDAvis inspection of Topic 8 of DMM 9-topic model

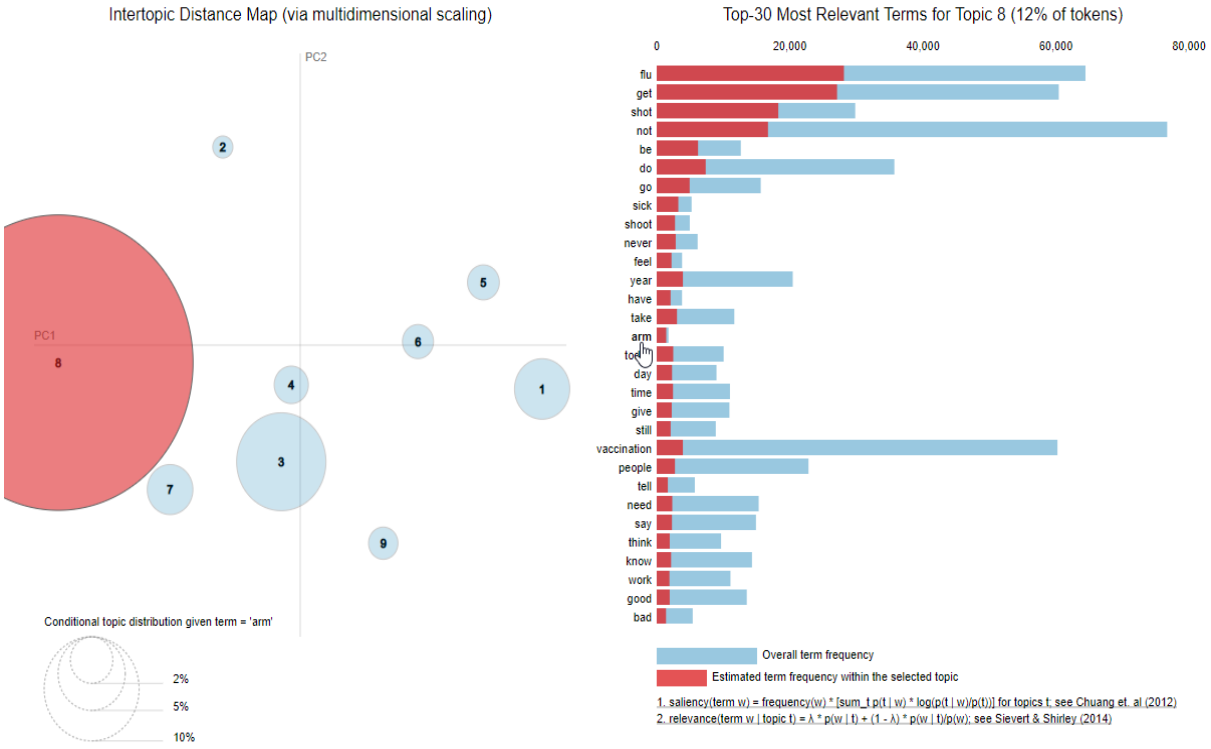


Figure 16: pyLDAvis inspection of "arm" in DMM 9-topic model

Gephi

The Gephi graph visualisation tool (*Gephi-The Open Graph Viz Platform*, n.d.) was also evaluated as an alternative to topic modelling, applying its implementation of the Louvain method (Blondel et al., 2008) for Community Detection. To prepare the data for Gephi a *graphml* file was created which defined nodes from every word in the corpus and edges up to 5 words away from each word within each document. Words that were closest to one another were given more weight on the edge compared to the words further away. The best result was that 31 communities were detected, but most of them were less than 1% of the records (in fact most around 0.01%), and only 3 major communities were detected. The most indicative words from the VAEM topic did not emerge as a distinct group. Although interesting to experiment with and to pick up relationships between words, graph visualisation of communities did not assist with detecting the kinds of distinct similarities obtained with topic modelling, and furthermore the graphing process was not helpful for the goal of clustering documents together for dimension reduction and labelling. Figure 17 shows a zoomed-in view of clusters of words per communities, one of them associates some of the VAEM keywords such as “get”, “flu” and “shot” but misses many of the other important words such as “arm”, “hurt” and “today”.

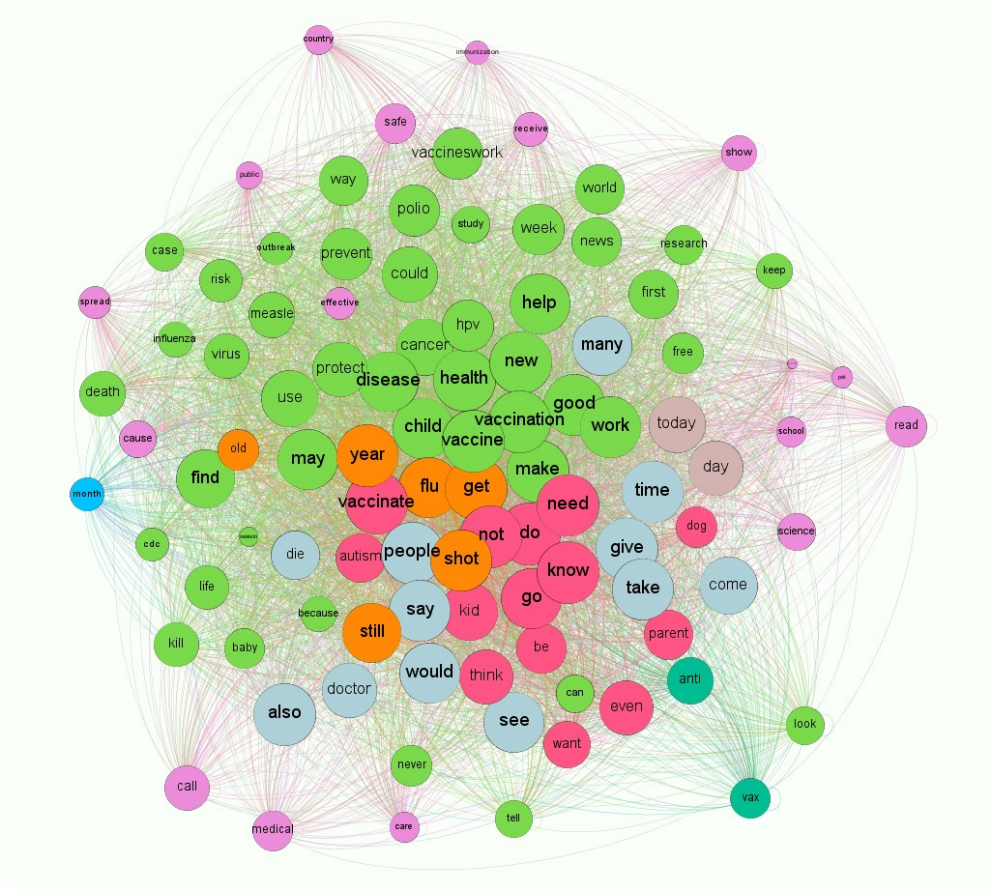


Figure 17: Gephi words clusters

5.9 Chapter 5 summary

An exploration of topic modelling leveraged a scoring technique that identified the most suitable topic models for identifying Twitter posts likely to contain vaccine adverse event mentions. The results of topic modelling were beyond the author's expectations, the data was reduced from an overwhelming cacophony of irrelevant posts to data that was fit for the goal of detecting vaccine adverse event mentions. The scoring technique was pivotal in assisting with making informed decisions about the best topic models, through two stages of topic modelling. The resulting filtered data was fit for proceeding to the next phase, which was to train classifiers to correctly identify the vaccine adverse event mentions from the now manageable dataset.

6 Classification

6.1 Chapter overview

The goal of using classification in this research is to further isolate VAEM in the data. The topic modelling step had reduced the input data to a manageable set of tweets that included almost all VAEM but also other similar posts; it was expected that classification would further refine this data to clearly extract most of the VAEM from the other posts.

Although the literature review into classifiers had shown that the most likely best classifiers would be deep neural networks (Deep Learning), it was decided that to obtain a thorough understanding of classifier performance a range of the traditional classifiers should be tested — these would provide a benchmark. As the classifier evaluation proceeded it was decided to also evaluate a rule-based approach to classification, to provide a baseline for all the classifiers. This is explained in Section 6.10.

6.2 Classifiers

Classification was conducted over a range of classifiers, starting with traditional classifiers, and including a rule-based classifier, then neural networks — these were either trained from scratch on the Twitter data, or used transfer learning on pre-trained models. The classifiers that were used are summarised in Table 25. Classifier parameters are specified in Appendix J.

Table 25: Classifiers

Traditional Classifiers	**Library / Github source**
Logistic Regression CV	sklearn.linear_model
Stochastic Gradient Descent Classifier	sklearn.linear_model
Linear Support Vector Machines	sklearn.svm
Random Forest Classifier	sklearn.ensemble
Extra Trees Classifier	sklearn.ensemble
Multinomial Naïve Bayes (Complement Naive Bayes on imbalanced datasets)	sklearn.naive_bayes
Naïve Bayes SVM (combined NB & 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

Neural Networks	Libraries / Github source
CNN, LSTM, BiLSTM, GRU, BiGRU, CNN-BiLSTM, CNN-BiGRU	Pytorch RaRe-Technologies/gensim Shawn1993/cnn-text-classification-pytorch bamtercelboo/cnn-lstm-bilstm-deepcnn-clstm-in-pytorch
RoBERTa, RoBERTa Large, BERT, XLNet, XLNet Large, XLM	Pytorch huggingface/transformers
ULMFiT	Pytorch, fastai

6.2.1 Calibrated Classifier Cross Validation

Initially, testing included an additional evaluation of the traditional classifiers when processed through the SKLearn.calibration library's CalibratedClassifierCV (CCV). This was for the purpose of obtaining probabilities for predictions and for the additional benefit of scoring with cross-validation (in technical terms: with CCV the classifiers were fitted and calibrated using cross validation and averaging). Probabilities were required for examining which words the classifiers favoured for differentiating between the labels; cross-validation gave a potentially more accurate estimate of the model's performance than any best single training run. Using CCV improved the F1-Scores of some models in an initial evaluation round and when using imbalanced data (Section 6.5). As models improved with better parameter tuning and with more training data then CCV did not improve scores and was no longer used.

6.2.2 Ensemble

An ensemble of 5 models was constructed: Naive Bayes SVM, Logistic Regression CV, SGD Classifier, Linear SVC, and Random Forest. Scoring these used a majority voting approach — whatever 3 or more models predicted was accepted — this had a correcting effect that compensated for variations in the models' performance over test datasets.

6.2.3 Neural network models

Training neural networks requires far greater computations, and these are most effectively carried out on specialized hardware, typically a Graphics Processing Unit (GPU). For this training, a NVIDIA GTX 1080 GPU was utilized, although when training the RoBERTa Large, XLNET Large and XLM models the 8GB memory on the GPU was insufficient, so training was performed on more powerful GPUs obtained on cloud-based dedicated deep learning

servers via Google Colab. Pytorch (Paszke et al., 2019) deep learning libraries were used throughout.

Neural networks typically use dense embeddings instead of sparse matrices to represent words to the models, these can start off in a random state or can be initialized using pre-trained embeddings. An initial test evaluated the models without any pre-initialized embeddings, but in every subsequent test Word2Vec embeddings were used where such embeddings were required (i.e., with the models that were trained from scratch). These clearly helped the models to learn more effectively. Models were trained on word sequences rather than at a character level.

The starting point for the exploration of the effectiveness of neural networks was with Convoluted Neural Networks (CNN). Various implementations of Recurrent Neural Networks were also evaluated, such as Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014) models. All these tests additionally assessed bi-directional versions of the RNN-based networks for their ability to better cope with arbitrary arrangements of text, and mostly these were found to be more effective. Benchmarks were established using the CNN, LSTM and GRU models trained on the first phase dataset, then compared those with results from training on the larger second phase dataset, to enable a comparison of the DL models effectiveness with varying amounts of data, and against the traditional classifiers.

6.2.4 Transfer learning

To look at some current state-of-the-art approaches to NLP, various Transformer-based models were then evaluated, using transfer learning. These were open-source HuggingFace (Wolf et al., 2019) sequence classification models that were based on pre-trained Transformer language models, e.g., *BertForSequenceClassification*. The HuggingFace implementations of Google's Bidirectional Encoder Representations (BERT); Facebook's Robustly Optimized BERT Pretraining Approach (RoBERTa); Google/CMU's XLNet; and Facebook's XLM were all evaluated. The XLM model is designed for language translation, it was included because HuggingFace had also made it available for binary classification and it was interesting to see how well it did. Though the XLM model performed well it was very demanding of GPU memory and took considerably longer than the others to train. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) (J. Lee et al., 2020), was considered — but its emphasis on biomedical terminology did not fit the prosaic and colloquial language of the tweets — there was almost no technical medical language used in them.

BERT was initially assessed against the smaller first phase dataset (see Section 6.2.5), and it did not perform that well. The much greater quantity of training data available in the second phase dataset was required to fairly evaluate the Transformers.

There are no validation scores listed in the results sections for these models since they are simply tuned to the existing data, and training and validation data was combined to tune them with. These models clearly performed better than most of the models trained from scratch, except the LSTM which outperformed both the BERT model and the XLNet base model.

Additionally, an experiment was conducted using Fast.ai's (Howard & Ruder, 2018) Universal Language Model Fine-Tuning (ULMFiT), an ASGD Weight-Dropped LSTM (AWD-LSTM) trained on Wikitext-103. It was fine-tuned with the Twitter data then applied to the classification task.

6.2.5 Evaluation measures

Overall performance was measured as a score based on Precision, Recall and the resulting F1-Score for the vaccine adverse event mention-containing records — this F1-Score is referred to as the VAEM F1-Score. Additionally, the same "Adjusted F1-Score" (Beta-F1 with beta of 1.3) that had been used during topic modelling was also evaluated, as the preference for identifying models that delivered a greater number of VAEM with the best possible precision still applied. During the training process the scoring was calculated against the validation data. The models' scores were then also compared against two hold-out test datasets.

6.3 Data preparation

For traditional classifiers, a "bag-of-words" approach was used, where documents are assessed as collections of words with no consideration to word relationships apart from what might be preserved by n-grams. Text was tokenized and vectorized as a sparse matrix of the document's words in relation to the entire corpus, which is the standard approach. A comparison between TF and TF-IDF vectorization (Section 2.5.2) was performed, utilizing the SKLearn CountVectorizer and TfidfTransformer (or alternatively the SKLearn TfidfVectorizer) libraries. The most appropriate text tokenization and vectorization approaches were determined for each model and dataset combination using grid or random searches. The text tokenization options that were evaluated consisted of stop-words elimination, the use of lower case, n-grams, and the exclusion of numbers.

Three forms of the text were evaluated, with examples in Table 26: the original text - which was used with some of the neural networks; a cleaned lower-case version, which expands

contractions, and was evaluated by all models; and lemmatized versions of the cleaned text, which was evaluated with the traditional classifiers. The reduced form of lemmatized text that worked well for topic models used only nouns, adjective, verbs, and adverbs — but this was not at all performant with the classifiers. Lemmatizing the complete cleaned text produced some mixed results (see Section 0), and this form is shown in the table.

Table 26: Text used in Classification

Original	It's still not too late to be getting you and your loved ones a flu shot! #vaccineswork
Cleaned	it is still not too late to be getting you and your loved ones a flu shot! vaccineswork
Lemmatized	it be still not too late to be get you and your love one a flu shot ! vaccineswork

The performance of the models was assessed by applying their best settings, as determined by grid searches, against both the cleaned text and the equivalent lemmatized text. The vector settings used are detailed in Appendix J.

As an alternative to a sparse matrix, with the standard models, experiments were also carried out with using the average of Word2Vec word embeddings per document; and additionally, with clusters of Word2Vec embeddings, see Section E.1 for a description. When evaluating neural network models, Word2Vec embeddings were used for the initial word embedding weights of the models, with a measurably positive effect compared to models with randomly (or default) initialized embeddings. These results are explained in the Deep Learning section.

6.4 Classification evaluation

Classification evaluation was conducted over two quantities of data. The first dataset (Section 4.6) consisted of the initial 6 months of data that was collected. The data had been used for training topic models - and then the records from the top 3 VAEM topics from the best second stage of topic model were exported and labelled for classification (Section 5.5). Experiments were initial conducted over all the labelled data, evaluating model performance with varying degrees of imbalanced data (see next section), then the data was balanced for the rest of the training. The second dataset (Section 4.7) included this data but added a further 6 months of data; it was obtained from the output of the top VAEM topic of the previously trained best *first* stage topic model (Section 5.3.4).

6.5 Initial experimentation with imbalanced datasets

The traditional classifiers were initially evaluated against four increasingly imbalanced datasets (Section 4.6.1), to gain some insight into how well they coped when trained and evaluated

using imbalanced data. The "Best" dataset contained just the records from the best topic of the second-stage topic model, Topic 8. The "Combined" dataset contained all of Topic 8, and just the VAEM and labelled Discussions records from Topic 9, plus the VAEM records from Topic 1, and was the most balanced dataset. The "Top Two Combined" dataset contained all of the records from topics 8 and 9, and the "All Combined" dataset contained all data from the three topics; these were progressively imbalanced. The datasets were split to apportion 30% of the data to a validation dataset, so validation sets had the same proportions of data imbalance as the training sets. Table 27 shows the dataset numbers used.

Table 27: Initial datasets training & validation splits

Dataset		Training	Validation	Total
Best	VAEM	929	375	1,304
	Non-VAEM	1,205	540	1,745
	Total	**2,134**	**915**	**3,049**
Combined	VAEM	1,222	518	1,740
	Non-VAEM	1,262	547	1,809
	Total	**2,484**	**1,065**	**3,549**
Top Two Combined	VAEM	1,056	409	1,465
	Non-VAEM	4,991	2,183	7,174
	Total	**6,047**	**2,592**	**8,639**
All Combined	VAEM	1,235	505	1,740
	Non-VAEM	11,728	5,051	16,779
	Total	**12,963**	**5,556**	**18,519**

The *class_weight* parameter of the models was adjusted to account for class imbalance, typically with an 0.75 weighting for the "Top Two Combined" dataset, and 0.85 or 0.9 for the "All Combined" dataset. Complement Naïve Bayes model was used instead of the standard Naïve Bayes model with the highly imbalanced datasets. The results are summarized in Table 28, which shows the best VAEM F1-Scores measured against the *validation data* of each dataset. In the table, the top 3 scores per dataset are shaded, with the top score having the darkest shade. If models performed best with TF-IDF vectors rather than just TF vectors this is indicated, as well as those that required CCV to achieve their best score per the dataset.

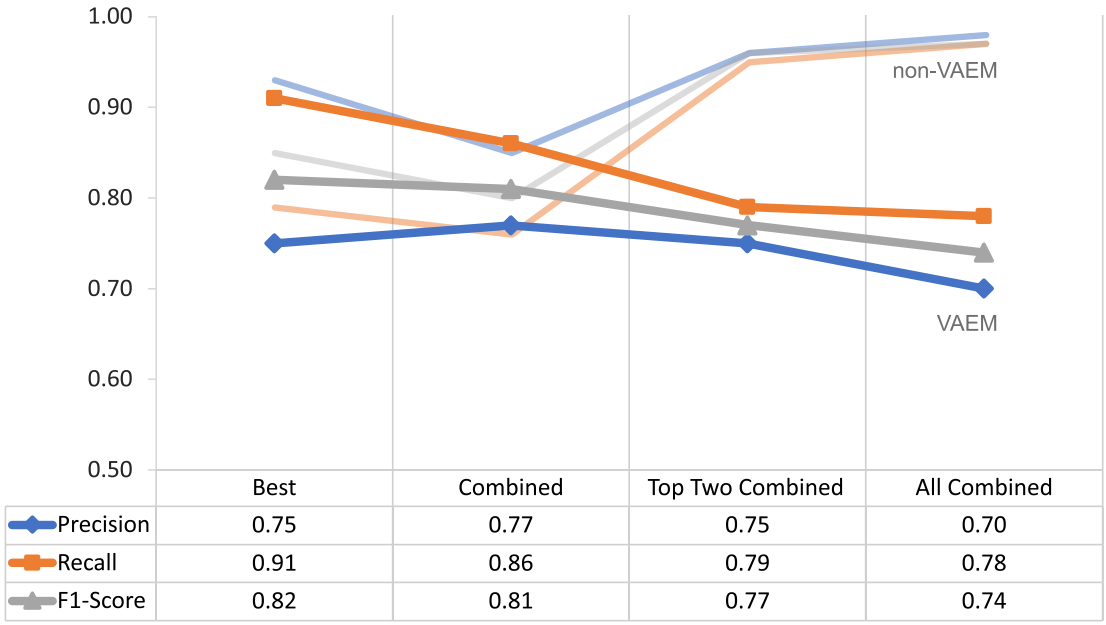
Table 28: Traditional classifiers - Initial F1-Scores

Model	Best	Combined	Top Two Combined	All Combined
Linear SVC	**0.823** ☼	0.811 ☼	**0.769** ☼	*0.740* ☼
Extra Trees	*0.820* ◆	**0.830**	0.735	0.708 ◆
Logistic Regression CV	0.818	0.806	0.756	**0.743**
Random Forest	0.815	*0.820*	0.731	0.695 ◆
Stochastic GD	0.808	0.812	*0.761* ☼	0.724
XG Boost	0.812 ☼	0.819 ☼	0.742	0.696
Naïve Bayes	0.813	0.808	0.732 ☼*	0.683 ☼*
Naïve Bayes SVM	0.791	0.789	0.721	0.685

* *Complement NB* ☼ TF-IDF ◆ CCV

Extra Trees outperformed the others on the balanced, "Combined" dataset but other models such as Linear SVC and Logistic Regression CV did well more generally, especially with the imbalanced data. The Extra Trees classifier benefitted slightly with additional cross validation folds (using CCV) on both the "Best" and "All Combined" datasets.

Models that performed well on the imbalanced data did so because they maintained a reasonably consistent balance of precision and recall in the VAEM class despite the penalty of training and validating with vastly imbalanced data. Figure 18 shows the VAEM scores of the Linear SVC model as it was trained and validated over the various datasets. The VAEM and non-VAEM groups of scores are labelled, and the VAEM values are shown in the table at the bottom of the chart.

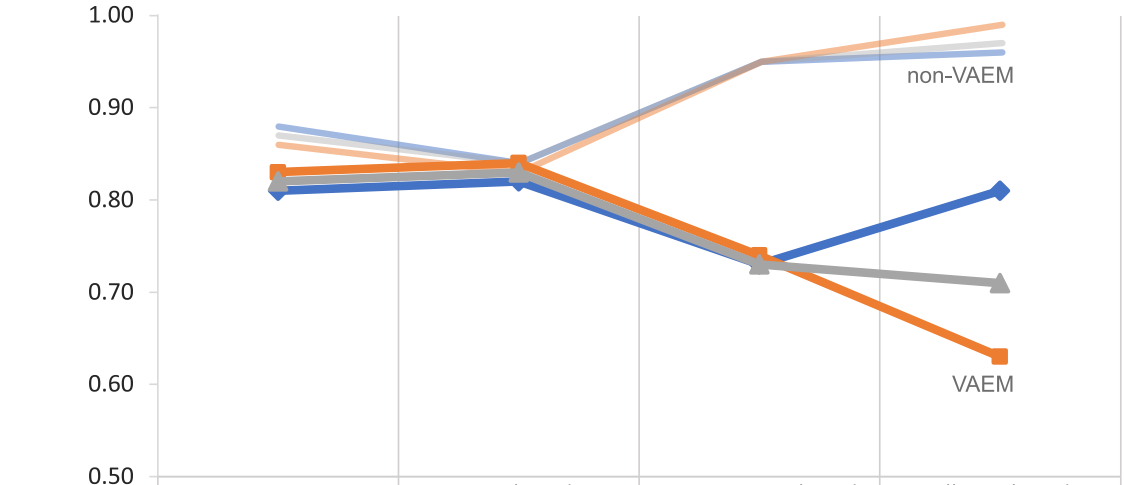


	Best	Combined	Top Two Combined	All Combined
Precision	0.75	0.77	0.75	0.70
Recall	0.91	0.86	0.79	0.78
F1-Score	0.82	0.81	0.77	0.74

Figure 18: Linear SVC - Scores over imbalanced datasets

The Linear SVC model's top VAEM F1-Score score was on the "Best" dataset, and this was mainly due to its high recall. Extremely high scores are obtained for the non-VAEM as the data becomes increasingly imbalanced, this is expected but has no benefit for identifying the VAEM, though it does illustrate why only the VAEM scores are considered when measuring model effectiveness.

The Extra Trees model had the highest VAEM F1-Score, but it did not perform well on the imbalanced datasets, its scores are shown in Figure 19. The model had very close relationships between the precision and recall until it got to the most imbalanced "All Combined" dataset, where recall suffered at the expense of precision. This is an effect of the model not coping with the imbalanced data. The model is assigning VAEM records to the non-VAEM side, which only slightly decreases precision on the non-VAEM side but very much increases precision on the VAEM side, at the expense of VAEM recall.



	Best	Combined	Top Two Combined	All Combined
Precision	0.81	0.82	0.73	0.81
Recall	0.83	0.84	0.74	0.63
F1-Score	0.82	0.83	0.73	0.71

Figure 19: Extra Trees - Scores over imbalanced datasets

The results of training on imbalanced data were expected and highlighted the difficulties of using such data. It was decided to use balanced datasets for training and to assess model performance on both imbalanced and balanced *test* data.

6.6 Experimentation with balanced training datasets

Non-VAEM records were removed to balance the data – see Section 4.7.2 for a description of the under-sampling approach used. The final balanced datasets used in both Phase One and Phase Two of the classifier's evaluation are described in sections 4.6.4 and 4.7.3. The Phase

One balanced main dataset contained 3,519 records, consisting of 1,722 VAEM records and 1,797 non-VAEM records, plus an imbalanced, holdout "Victorian" 614 records test dataset which consisted of 90 VAEM and 524 non-VAEM - see Table 7.

The Phase Two balanced dataset included these but added additional records, resulting in 9,564 VAEM and 9,685 non-VAEM records in the 19,249 records main dataset used for training and validation, plus a balanced 828 records test dataset made up of 431 VAEM and 397 non-VAEM — see Table 9. The models were assessed against both test datasets. The distribution of these records into training, validation and test datasets is depicted in Table 29.

Table 29: Final datasets training, validation, and test splits

Dataset		Training	Validation	Test	Total
Phase One	VAEM	1,291	431	90	1,812
	Non-VAEM	1,348	449	524	2,321
	Total	**2,639**	**880**	**614**	**4,133**
			3,519		

Dataset		Training	Validation	Test	Total
Phase Two	VAEM	7,512	2,052	431	9,995
	Non-VAEM	7,545	2,140	397	10,082
	Total	**15,057**	**4,192**	**828**	**20,077**
			19,249		

6.7 Phase One classifiers results

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 Table 30.

Table 30: 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
Rule-Based	0.745	0.656	-	-
Naïve Bayes	0.798	0.605	0.799	0.756

The table 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 three best F1 scores on each of the datasets per classifier category are shaded, with the best score having the darkest shade.

The Ensemble model is shown in the middle of the table, which was scored based on a maximum voting of the predictions of 5 standard 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.

Experiments with a lemmatized version of the cleaned text showed that some of the models (especially the Logistic Regression CV, Linear SVC, and Stochastic Gradient Descent models) increased their validation F1-Scores when training on lemmatized texts, by as much as 0.2. However, all the models scored worse, in some cases worse by even more than 0.2, when assessed against lemmatized versions of both the imbalanced and balanced *test* data. They seemed to be over-fitting on the lemmatized training data and were not as generalizable on unseen data — therefore lemmatization was not used.

A baseline rule-based classifier was constructed during the first classification phase - this is described in Section 6.10. Its performance was considered as a baseline for the classifiers, and its best score on the Imbalanced test scores has been included on the bottom of the table. It was not re-evaluated in Phase Two. The Phase One evaluation process also included assessing alternative word embedding vectorization approaches against the test dataset, but these approaches were not based on the "off the shelf" vectorizing techniques for standard classification. One used the average of Word2Vec word embeddings per document, and the other used "centroids" - clusters of Word2Vec embeddings. These resulted in mostly poorer scores than the sparse matrix approach — see Appendix E.

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.

A very best F1-Score of 0.805 on the imbalanced test set was obtained with a "manually-tuned" CNN, which was a one-off result from a test of manually decreasing learning rates in small steps. This result was not able to be reproduced by using an automated learning rate adjustment scheme, so it has not been included in the chart - but obtaining this score indicated the possibility of an optimum result with a careful training regime. By comparison, 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.

6.8 Phase Two classifiers results

As depicted in Table 29, training in the second phase of classification used five times as much training data, by combining the 3,519 training records from the first phase with another 15,730 records, resulting in a total of 19,249 training records. Phase Two also introduced the Phase-Two Test dataset of 828 records. The greater amount of data allowed a proper assessment of neural networks and meant that additional Transformer models and the ULMFit model could be assessed. The scores for all models were considerably better in the second round of testing, especially over the imbalanced test dataset, and different best models emerged, see Table 31.

Table 31: 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%

Several models' names have been abbreviated, compared with the previous table: Logistic Regression CV is "Logistic RCV", Stochastic Gradient Descent is "Stochastic GD", and Naïve Bayes SVM is "NB SVM". The best individual traditional models were the Logistic Regression CV and Stochastic Gradient Descent classifiers. The Ensemble was the best performing

traditional classifier across all test datasets, but only by a marginal 0.001 on the Balanced test set and 0.004 on the Combined test set, compared to the Logistic Regression CV model.

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. The only Transformer trained in Phase One was the BERT model, so it is the only Transformer with comparative scores. Many of the score increases are very large, as much as 26.7% for the Stochastic Gradient Descent model on the Imbalanced Test data, and 10.3% for the LSTM on the Combined Test data.

The two arrows on the left of the table indicate the approximate positions of the maximum Imbalanced Test F1-Scores from the Phase One traditional and deep learning models, which were 0.767 for the Naïve Bayes SVM model, and 0.807 for the BiLSTM model. The previous maximum Combined Test F1-Scores, which were 0.825 for the CNN-BiGRU and 0.801 for the Ensemble, are both below the Phase Two Combined scores so cannot be similarly represented. As previously explained, validation scores are not applicable for the Transformer models, but validation F1 scores have increased by around 5% where numbers are available.

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.

One of the most noteworthy effects of having more data is that the previously strong combinations of CNN and bi-directional 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.

The ULMFiT model scored unexpectantly poorly, but it is likely that its mediocre performance is due to the author not knowing how to best tune it and that the score is misrepresentative of the classifier's capability. The model had an excellent recall (behind only the best Transformer models) but was overly sensitive to nuances in language and generated too many false positives — for instance in testing predictions it would tip an otherwise non-VAEM prediction over to a VAEM prediction just by the inclusion of one of the translated "sad" emoticons. Therefore, although the result is listed in Table 31, the model has not been included in the analysis in Section 6.9.

In conclusion, having more data has improved the results. With the deep learning models trained from scratch there was an average of 0.1 F1-Score improvement over the previous result on the Imbalanced test data, from an average F1-Score of 0.733 to an average of 0.833. Most tellingly the Transformer-based BERT model improved by almost 13% with the addition of the extra data and so performed as expected, with an F1-Score of 0.8634. Finally, having more data enabled an evaluation of the powerful RoBERTa Large model, which achieved an F1-Score of 0.919 on the Imbalanced Test data.

6.9 Classifier performance over the two training phases

The results depicted in the F1-Scores tables above show an expected effect, that after training with more data, almost all the models improved considerably, and that the scores were more consistent over the test datasets. So far, F1-scores have been used to evaluate and compare the models. However, as with the topic modelling goals of obtaining the best recall with an optimum possible precision, the same F1-Beta score (with a beta of 1.3) was also evaluated with the classifiers. The difference between F1 and F1-Beta, together with an analysis of the constituent precision and recall, helps to identify the top classifiers as well as highlighting classifier differences in handling the test datasets. For example, Table 32 shows the F1 and F1-Beta scores and the constituent measures for the RoBERTa Large model over the test data.

Table 32: RoBERTa Large F1 scores and measures

RoBERTa Large	TP	TN	FP	FN	Precision	Recall	F1	F1-Beta
Balanced Test	409	336	61	22	0.870	0.949	0.908	0.918
Imbalanced Test	85	514	10	5	0.895	0.944	0.919	0.925
Combined Test	**494**	**850**	**71**	**27**	**0.874**	**0.948**	**0.910**	**0.919**

6.9.1 Imbalanced Test data with Phase-One models

The Phase-One Imbalanced Test dataset was used throughout as a standard because it represented a real scenario of tweets that had been first identified as belonging to the author's geographical region, and its proportion of VAEM to non-VAEM was not balanced. During the initial Phase-One testing, the author 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 standard models, but it also eliminated true positives and so had the poorest recall. However, it was awarded the highest F1 Score among the standard 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.

When testing with an F1-Beta score it was observed that the Naïve Bayes SVM F1-Beta score was closer to the middle of the scoring range, and that BERT, with its relatively higher recall, was promoted to second place — see Figure 20.

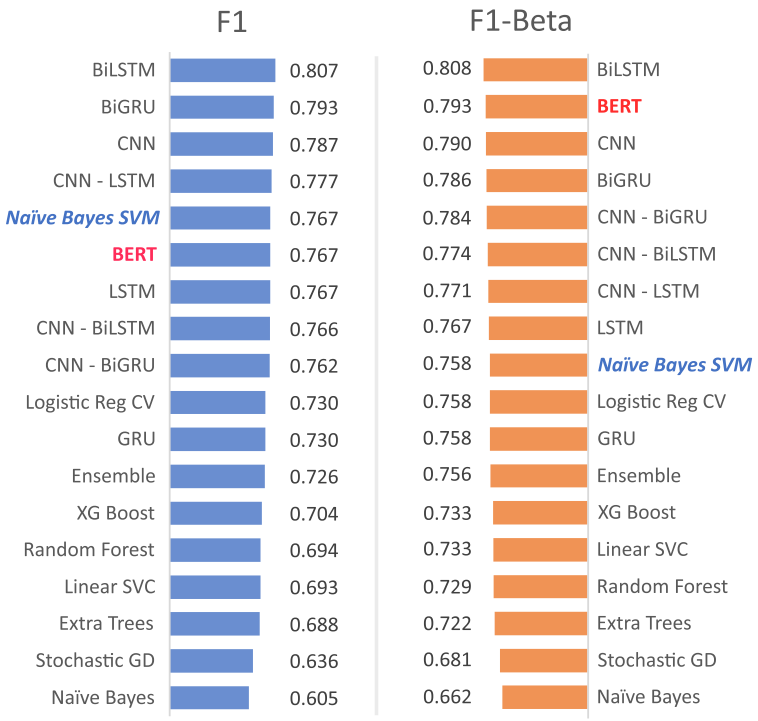


Figure 20: Phase-One models and Imbalanced Test data - F1 vs F1 Beta

The relationships between F1 and F1-Beta in the Phase-One models on the Imbalanced test data, together with the variations of precision and recall that help to explain the model's performance, are depicted in Figure 21. 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 standard models 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.

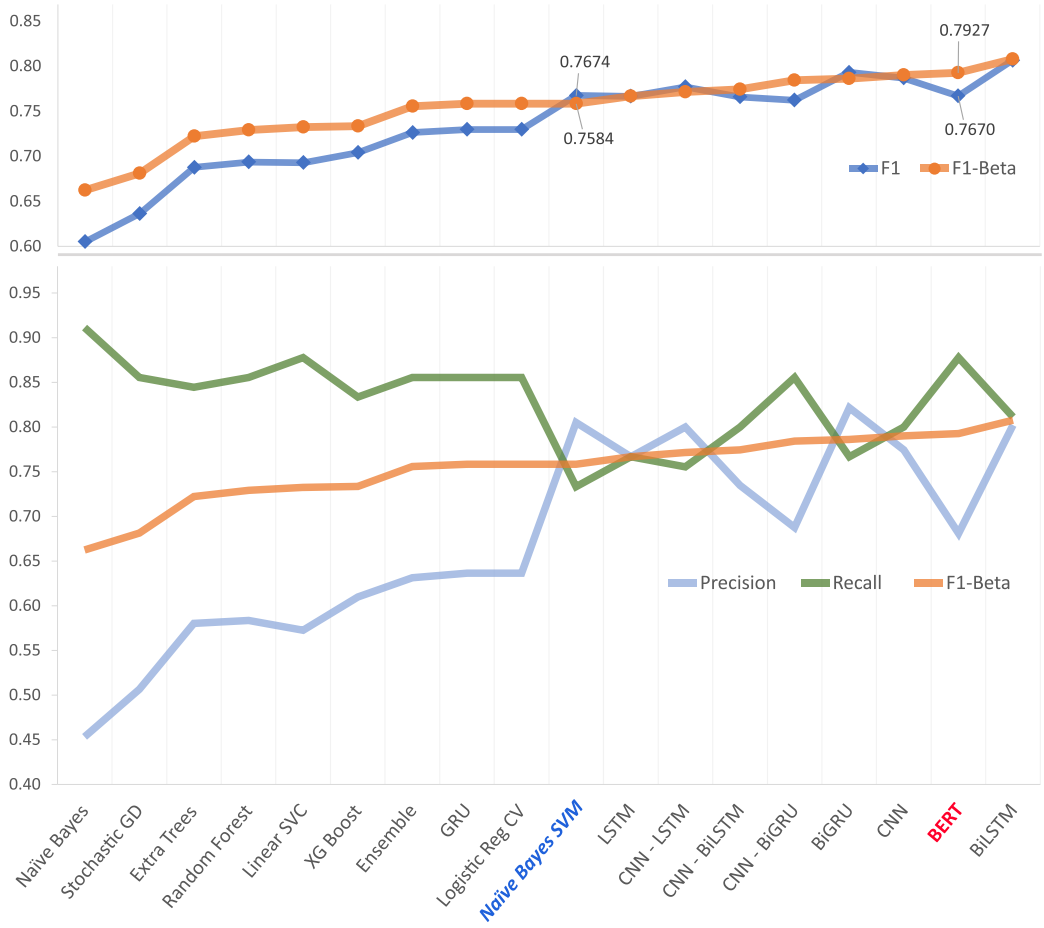


Figure 21: Phase-One models on Imbalanced Test data - F1 Scores & measures

6.9.2 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 five-fold increase in data available in the second phase of training - where the training data increased from 3,519 records to 19,249. The Phase-Two classifiers included five new Transformer models. See Figure 22 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 trained 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 standard models, and effectively coping with the imbalanced data — the "Balanced Test data with Phase-Two models" section (6.9.4) 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.

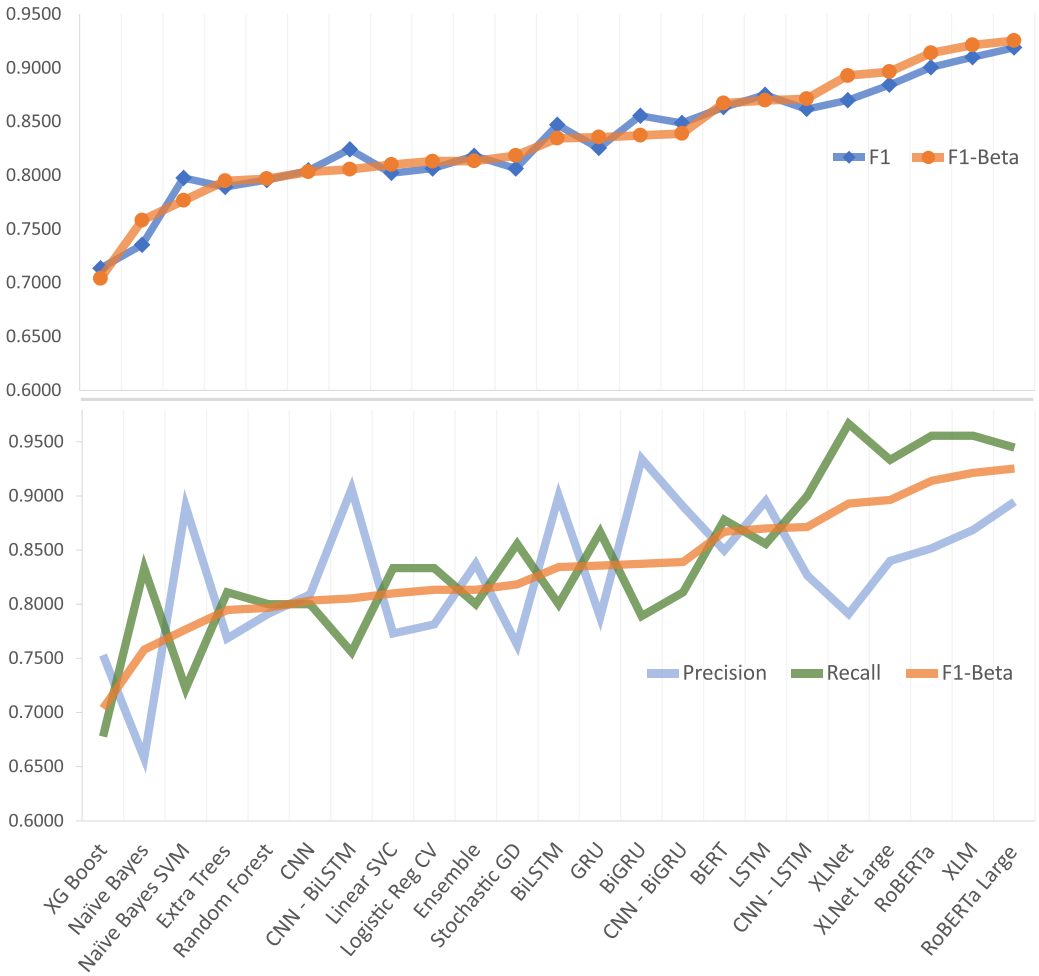


Figure 22: Phase-Two models on Imbalanced Test data - F1 Scores & measures

6.9.3 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 Figure 23. 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 — a very high recall was weighted by the F1-Beta calculation to offset a poor precision and give the model 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 standard 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 standard models and neural networks, several standard classifiers performed better than some of the neural networks.

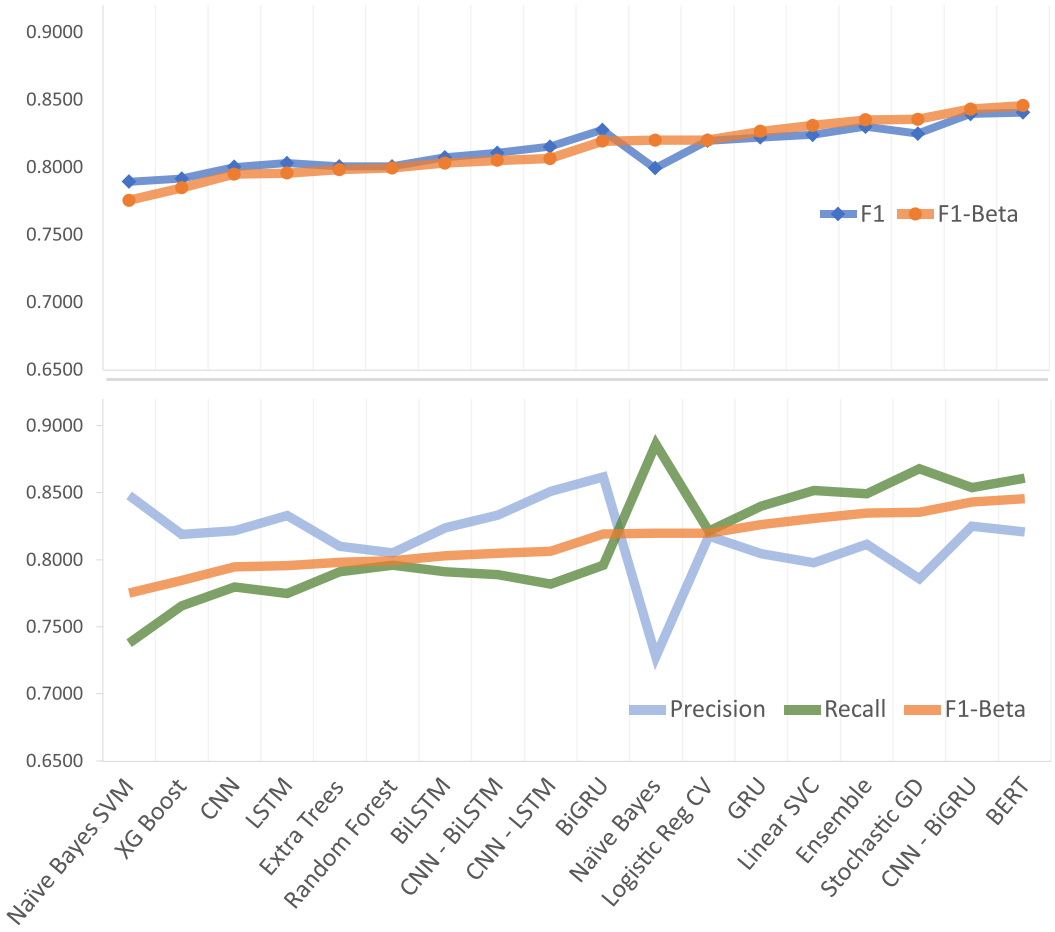


Figure 23: Phase-One models on Balanced Test data - F1 Scores and their measures

6.9.4 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 Figure 24.

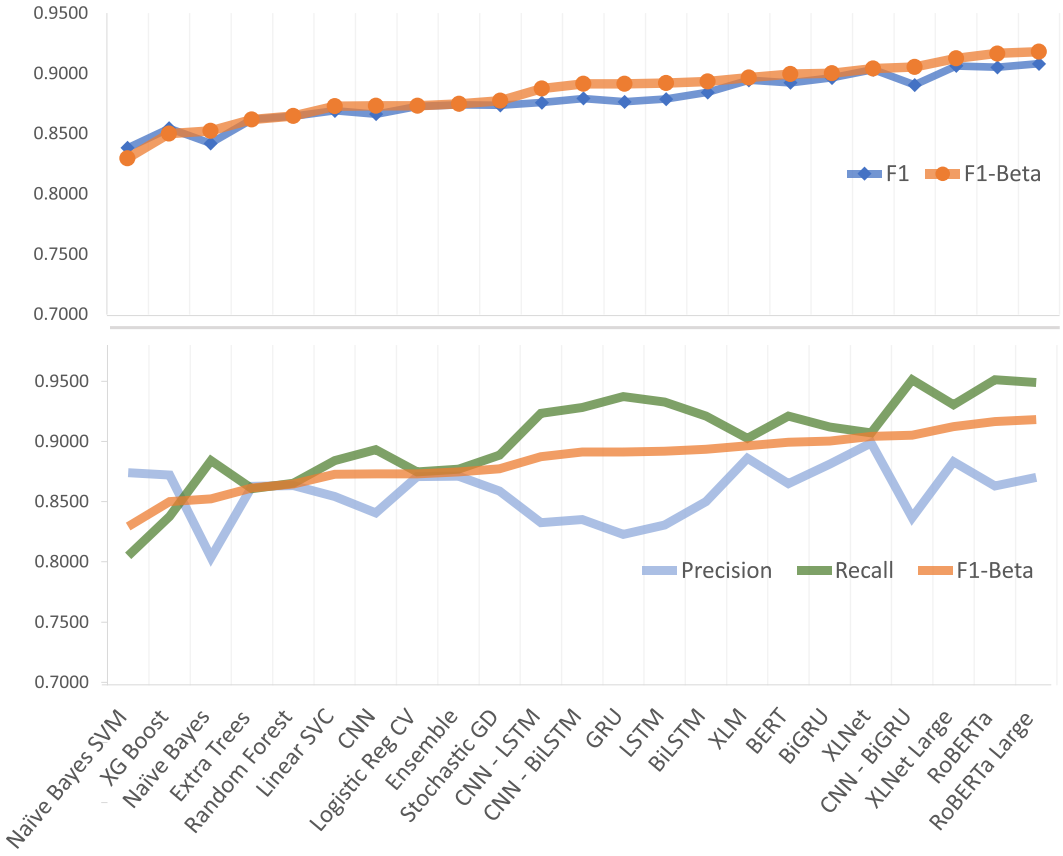


Figure 24: Phase-Two models on Balanced Test data - F1 Scores and their measures

The Phase-Two results show an even closer relationship between the F1 and F1-Beta scores, with a clear progression in scores from the standard models to the neural network-based models — there are no longer standard 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 the author concludes that training on the large dataset has achieved an optimal result.

6.9.5 Phase-Two models vs Phase-One models

Figure 25 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.

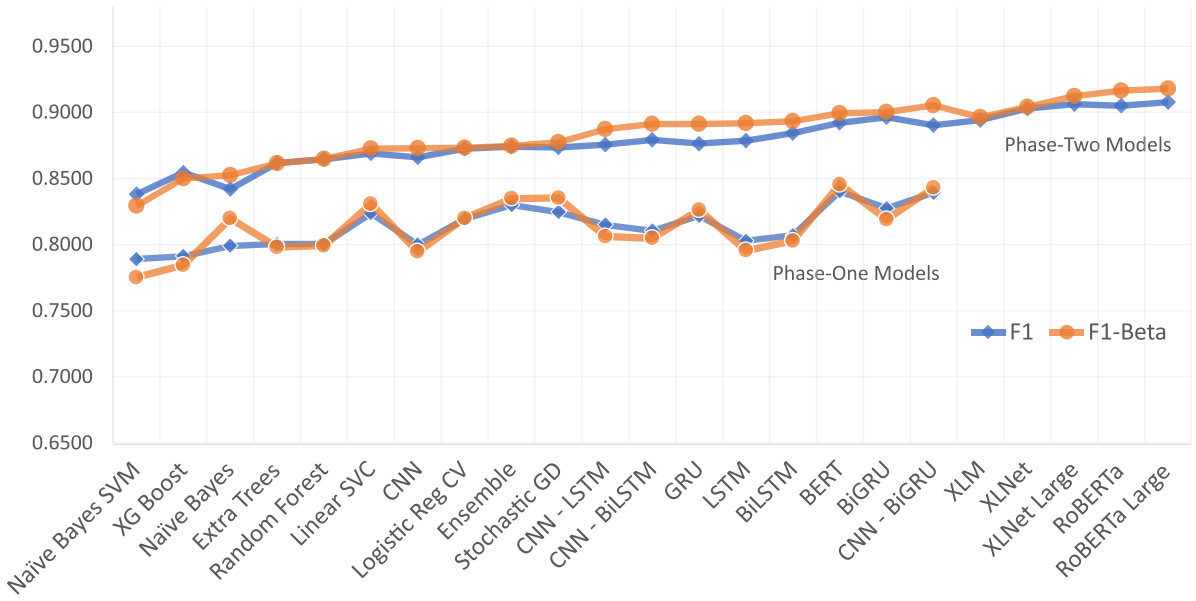


Figure 25: Phase-Two vs Phase-One models - F1 scores on Balanced Test data

To conclude this section, the descriptions of classification results and the analysis presented above highlighted the need for sufficient data to train classifiers and demonstrated the effect of imbalanced test data when classifiers are favouring the dominant, negative, class. The analysis 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. The section also demonstrated that the most powerful classifiers were able to cope with imbalanced data and were consistent in their performance.

The F1 scores of the Roberta Large model are exemplary — the almost 0.92 F1-Score on the Imbalanced test set is more than 10% better than the best traditional classifiers, which was the Ensemble. In Phase Two the Ensemble improved by 11.3% over its score from Phase One, but since the RoBERTa Large improved over this score by 10%, F1-Scores improved by 21% through the combined effect of increasing the training data and utilizing transfer learning with state-of-the-art Transformer models. The results have established that classification can very effectively isolate VAEM from the incoming topic model data.

6.10 Baseline rule-based classification technique

The most significant VAEM words for the traditional classifiers tended to focus on complaints of side-effects such as pain and sickness arising from a recent vaccination. The patterns were repetitive enough to suggest that a rule-based approach could be designed as a baseline technique. Therefore, it was decided to utilize the most-used words in VAEM-containing posts

as a target for scoring *document similarity* to those words. At the same time, the most-used VAEM words were also present in non-VAEM posts, so a rule-based system should not just classify every post as VAEM on finding common VAEM words. Instead, it should try to compensate with some awareness how these VAEM words were used, or to take account of non-VAEM words that might be used to counterbalance the tendency to make false positive predictions.

The target VAEM words were those which were observed to be present in most VAEM-containing documents, and in significantly greater proportion to their presence in non-VAEM-containing documents. The opposite condition was also applied to find non-VAEM target words. Some words were observably common, for instance “arm” and “hurts” in VAEM documents. Others were chosen based on their presence in the same cluster as the most significant words when k-means clustering (Section 2.6) was performed on the entire data. Appendix Section D.3 contains a description of how these clusters were constructed.

VAEM target words included “arm”, “hurts”, “got” and “yesterday”. Non-VAEM target words included “get”, “vaccine”, “needles”, “test” and “jesus” (often in the context of “Jesus is my flu shot”, a popular meme in the texts). That is, it was observed that a lot of the discussions in VAEM-containing documents were about having painful arms due to getting a vaccination sometime in the previous day or days, and the term “vaccine” was hardly used. Non-VAEM-containing documents, however, were more likely to discuss vaccines and encouragements to “get” them, to include information related to “test”, to complain about the fear of getting “needle(s)”, or to avoid them altogether because of beliefs (e.g., due to Jesus’s protection).

Similarity scoring uses the built-in functionality of Word2Vec — words have a value in the vector space of the Word2Vec model, and similar words have a similar value — the similarity can be measured via the Word2Vec function *most_similar*. For the underlying Word2Vec embeddings, three models were assessed: (1) a model with word embeddings on all the words in the underlying data; (2) one with embeddings only on words greater than one character in length; and (3) a smaller model with embeddings on words greater than one character in length *and* with all the common English stop words removed. The smaller model gave the best results, the similarity scores between the words were stronger with the removal of the stop words “noise”.

For the rule-based approach, words in a document were scored based on their similarity to the VAEM and non-VAEM target words, by using the Word2Vec *most_similar* function applied to the Word2Vec model. If a word being processed was *not* found in the Word2Vec

model, then it could not be scored, unless it happened to be one of the target words. For instance, the word "get" is quite highly indicative of a non-VAEM and so could be used as a target non-VAEM word, but "get" is also a stop word and so did not appear in the smaller Word2Vec model. Nevertheless, if "get" was used as a target word and "get" is encountered in a document then obviously they perfectly match and so the "get" must be assigned a similarity score of 1, despite it not being found in the Word2Vec model.

Scoring a document would firstly consist of assessing every word in the document to get its similarity score to each of the VAEM target words (e.g., target words "arm" and "hurts"). Secondly, those scores would be sorted to get the top *n* scores, where *n* was a parameter (e.g., 5), and the top *n* scores would be added for that word. Thirdly, the top *n* scores in the document would be determined by sorting the scores, and then were added to get a total document score. The words that earned the top *n* scores would be preserved, so the output would be a score, and the most significant *n* words in the document, in descending order of their similarity score.

Non-VAEM target words were optional, but if specified then scores against them would be kept separately, so a document would have a VAEM score and words, and optionally a non-VAEM score and words. Experimentation indicated that a parameter of five worked best for the number of word similarities to target words.

To avoid inflated scores due to repetitive use of high-scoring words, the words would be counted as processed, and the count would be used as a denominator to adjust the score. If a word appeared for the first time its count would be 1 and with 1 as the denominator it would get a full score. Every subsequent appearance of the word would increase its count and denominator - so its subsequent scores would be progressively penalised.

There were two approaches for obtaining the similarity scores: One was to ensure that only certain words would get scored by utilizing limited dictionaries of potential words (e.g., the top 200 similar words to the chosen VAEM target words, derived by k-means clustering of similarity scores). The advantage of this approach was that only some words would ever be scored, they had to already exist as known similar words — therefore having a document score above a certain threshold was a fairly reliable indicator of it being a VAEM document. The disadvantage was that a document could end up with no score at all if none of its words were found in the dictionary. An example of scoring the sentence "*my shoulder is tender and sore and I feel sick*" with this approach is shown in Table 33. The target words are "arm", "hurts", "pain", "reaction", "fever" and "sick". Some scores are zero because the limited dictionary has removed any words that were not in the top 200 similar words to the target words, despite any similarity they had.

Table 33: Rule-based model - Limited dictionary of similar words

Top Words - Limited Similarity Scores									
	Word	arm	hurts	pain	reaction	fever	sick	**Total**	**Top 5**
1	sore	0.872	0.813	0.655	0.000	0.000	0.000	**2.340**	
2	shoulder	0.792	0.672	0.631	0.000	0.000	0.000	**2.095**	
3	feel	0.638	0.643	0.000	0.000	0.000	0.000	**1.281**	
4	sick	0.000	0.000	0.000	0.000	0.000	1.000	**1.000**	
5	my	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	**6.716**
6	is	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
7	tender	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
8	and	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
9	and	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
10	i	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

The other approach bypassed the use of dictionaries and always scored words using the Word2Vec *most_similar* method, so it would calculate a similarity no matter how different a word might be from the target. Scoring all words ensured that all documents would be scored, but it also meant that it was more complex to decide whether the score could be considered indicative of a VAEM document. This approach needed to account for non-VAEM scores and deciding that a document was VAEM based on the difference between the two scores.

Table 34 illustrates scoring where all words are scored because scoring has not been restricted to words available only in limited dictionaries.

Table 34: Rule-based model - All similarity scores

Top Words - All Similarity Scores									
	Word	arm	hurts	pain	reaction	fever	sick	**Total**	**Top 5**
1	sore	0.872	0.813	0.655	0.462	*0.408*	0.588	**3.390**	
2	feel	0.638	0.643	0.571	0.458	*0.274*	0.591	**2.901**	
3	shoulder	0.793	0.672	0.631	0.478	*0.245*	0.321	**2.895**	
4	sick	0.460	0.462	0.423	0.364	*0.343*	1.000	**2.709**	
5	my	0.520	0.563	0.435	0.412	*0.220*	0.479	**2.409**	**14.304**
6	tender	0.497	0.509	0.421	0.320	0.429	*0.264*	2.176	
7	and	*0.231*	0.268	0.332	0.309	0.258	0.356	1.523	
8	is	0.261	0.370	0.205	0.373	*0.175*	0.334	1.543	
9	and	*0.116*	0.134	0.166	0.155	0.129	0.178	0.762	
10	i	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

Note that only the top 5 scores per word are taken to score the word, and only the top 5 scores are taken to score the document. Any scores earned outside of the top 5 limits are not utilized and are depicted with a lighter grey colour in the table. In this case the word “fever”

did not contribute anything to the individual word scores in the top 5 words, as all the other words scored higher. Also note that the word “and” earns only half the score for the second instance of the word, which illustrates the penalty assigned to repeated words. The word “I” does not earn any points at all, as it was excluded from the Word2Vec model.

Three document scoring strategies were devised to make use of these scoring approaches. They were evaluated for F1-Scores using both validation data and the Imbalanced test data (the only test data available at the time of designing this approach).

A first document scoring strategy was to just score VAEM words and to limit words to those that could be found in a dictionary of most similar words. When scoring this way, a document would be considered as a VAEM if its score were greater than zero. When evaluating the scoring technique against the document labels and specifying the VAEM target words “arm” and “pain”, an F1-Score of 0.771 was obtained on the validation data. An F1-Score of 0.627 and an F1-Beta score of 0.644 were obtained on the test data.

The second document scoring strategy was to again use dictionaries to limit the words that could be scored, but to also score non-VAEM words. A dictionary size of 200 words was found to be optimal for VAEM words and 400 for non-VAEM words. This approach assigned a VAEM label to a document when the VAEM score exceeded zero, but only if the non-VAEM score was no greater than 1. Using a pattern of “arm” and “hurts” for VAEM and “not” and “get” for non-VAEM, an F1-Score of 0.745 was obtained with the validation data and 0.656 on the test data. The F1-Beta score on the test data was 0.664. This model is depicted in the Phase One model scores in Table 30.

The third scoring strategy used an “always score” approach. It would score all words in the document, and score a document as being a VAEM post if the VAEM-related score exceeded the non-VAEM-related score, but by an added threshold. For instance, if a threshold value was 2.0 it meant that the VAEM-related score needed to be at least 2 points higher than the non-VAEM-related score for a document to be considered a VAEM. When specifying VAEM words of “arm”, “hurts”, “got” and “yesterday” and non-VAEM words of “not”, “get”, “needles”, “test”, and “jesus” and using a threshold of 2.0 the resulting validation F1-Score on the example document was 0.7631. A higher threshold of 2.9 was required with the same words on the test data to correctly classify the document, resulting in an F1-Score of 0.645, and an F1-Beta also of 0.644.

Table 35 depicts the third scoring strategy, with examples of tweets, their VAEM scores and top words, and the rule used for that top word (V score, V top word, V top rule); their non-VAEM counterparts (NV score, NV top word, NV top rule); and the difference between the

scores (Diff.). The VAEM rule words were “arm”, “hurts”, “got”, and “yesterday”, the non-VAEM rule words were “not”, “get”, “needles”, “test” and “jesus”. The decision about whether a tweet should be classified as a VAEM is based on the threshold that is applied to the differences in their scores. If a threshold of 2.5 is applied, then all records with a difference of 2.5 and above would be correctly classified, but if the threshold is 2.0 then the third record would be incorrectly classified as VAEM, as the 2.219 difference in the score exceeds the threshold of 2.0.

Table 35: Examples of manual scores

Tweet	Class	V score	V top word	V top target	NV score	NV top word	NV top target	Diff.
i got my flu shot today and my arm is so sore i cant do anything...soooo anyone wanna cuddle?	VAEM	13.895	got	got	9.184	soooo	needles	4.711
i also have to get my vaccine shots today <annoyed> but at least i'm not scared or terrified of needles i just dislike them	Non-VAEM	10.352	my	got	10.366	needles	needles	-0.014
i'm terrified of getting shots/vaccinations or having people take blood because i'm always scared they will snap the needle off in my arm.	Non-VAEM	11.413	arm	arm	9.194	scared	needles	2.219
feeling horrible today after getting the flu shot so annoying	VAEM	11.124	feeling	yesterday	8.459	annoying	not	2.665
flu vax- thoughts? i'm very much for.	Non-VAEM	3.835	flu	got	3.876	flu	not	-0.041

These techniques provided a baseline rule-based approach to measure classifiers against. The only classifiers which performed worse than the rule-based approach were the Naïve Bayes SVM and Naïve Bayes models in Phase One of the classification, with F1-Scores of 0.5970 and 0.5292 on the Imbalanced Victorian dataset. All the classifiers surpassed the rule-based approach in Phase Two of the classification, which involved re-training with more data. The various scoring strategies used in the rule-based scoring approach are relatively successful depending on the dataset and the application of a lot of manual tuning to get the best from them — but the approach is not robust enough to adopt as an alternative to any of the machine learning models. Nevertheless, the techniques suggested that extra embeddings-based features could be derived that might be applicable to improve a classifier’s performance, which is explored in Appendix E.

6.11 Chapter 6 summary

This chapter explored the capacities of classifiers in terms of small and larger dataset sizes, resulting in two phases of classification evaluation. This was important, as data availability is a key consideration when choosing classifiers and it is not always possible to find the data that is required to properly train the state-of-the-art deep learning classifiers. By providing the performance detail of a range of classifiers in relation to data availability, the aim was to help inform fellow researchers to assist their choices when collecting data and choosing classifiers. Section 7.3 of the Evaluation chapter further compares these results using charts.

7 Evaluation

7.1 Chapter overview

This chapter contains detailed evaluations of the performance of the topic modelling approach and classifiers that were used in this research. The analysis confirms the effectiveness of the techniques for identifying posts containing vaccine adverse event mentions, by measuring how the Twitter posts containing VAEM have been successively concentrated by applying these techniques. The chapter also clarifies that classifier effectiveness is relative to data availability.

7.2 Evaluating topic model effectiveness

Topic model effectiveness measures the capacity of the topic models to identify VAEM in the texts by bringing them together into one or two topics, so that exporting documents from only these topics can be used as a filtering mechanism. The most effective topic (the top topic) would be one that contains all or most of the VAEM without containing very many other documents. The following sections take the approach of counting proportions of VAEM in the top topics vs. VAEM in all topics (like a recall score), and VAEM vs. all other labels in a top topic (like precision). In section 7.2.1 the discussion is based on *labelled samples*, and in section 7.2.2 it is based on the fully *labelled classification data.*

7.2.1 Verifying topic models with samples

As previously discussed, Topic 13 of the 14-topic stage-one DMM model had been identified as the best topic, by the scoring system that utilized a small set of 1,400 labelled tweets. The Topic 13 tweets were subsequently labelled for classification purposes as either VAEM or not and used for the ongoing work. The remaining thirteen topics of the 14-topic model were not labelled and were not henceforth used, and beyond the insight gained by the topic model scoring system, it was unknown whether these might actually contain VAEM.

Verifying the stage-one 14-topic model

To verify that it had been justified to discard topics other than Topic 13 as unlikely to contain VAEM, ten thousand samples were taken of the *total* data of the 14-topic model (i.e., inclusive of the already labelled Topic 13), and the as-yet unlabelled tweets were labelled - as either VAEM or not. The result was 10 samples, each consisting of 10 randomly sampled groups of 100 records and combined into 1,000-record samples, for a total of 10,000 sampled records. Only one VAEM record emerged in another topic apart from Topic 13, and this is depicted in

Figure 26. For simplicity only topic numbers containing VAEM are assigned a separate bar in the chart, all other topics are combined into the third "All Others" bar. Within topics 13 and 1 ("Thirteen" and "One" on the chart) the VAEM containing tweets are depicted by the darker portions on the right of each bar, the lighter portion to the left is non-VAEM. These are counted, for instance in Topic 13 there are 127 VAEM and 1,564 non-VAEM, a total of 1,691. The percentage label to the right of the bar for Topic 13 indicates the proportion of all VAEM in the topic, which is 99.2%. That is, 127 of the total 128 VAEM in the samples are found in Topic 13, and therefore only 0.8% of VAEM are found in other topics. Topic 13 consists of only 16.9% of the total sampled data, which is consistent with the filtering effect observed during topic modelling.

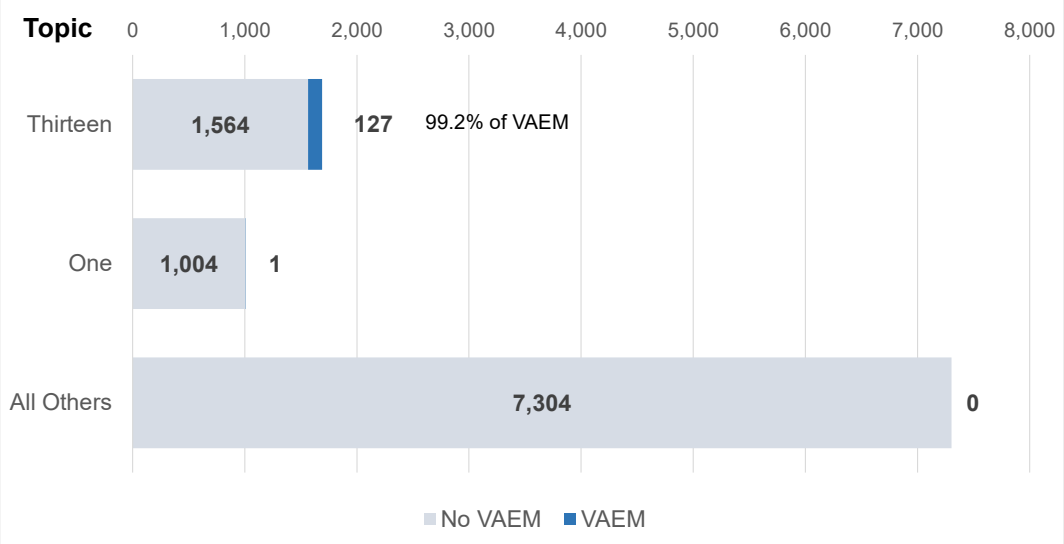


Figure 26: Sample Distribution of VAEM Stage One

Five additional 1,000 sized samples were taken of topics *excluding* Topic 13, only 2 VAEM were found in the 5,000 samples, which was 0.04% of the data. These numbers verify the effectiveness of Topic 13 of the 14-topic DMM model for filtering vaccine adverse event mentions. To be consistent with the samples and allowing for error it is estimated that *99% of the vaccine adverse event mention data* is being included in records collected by Topic 13 of the DMM 14-topic model.

Visualizing filtering with the taxonomy topics

Figure 27 illustrates the change of document topics as the data is progressively filtered over two stages of topic modelling to select the best VAEM topic. Samples of records were assigned the subjects of the *taxonomy*, which corresponds to the dominant topic of the tweet. The chart

compares the proportions of taxonomy topics found in the Stage One, Stage Two, and Topic 8 samples. The bars depict the spread of taxonomy topics in the data that each stage handles. The "Sample = Stage One" section represents the taxonomy topics of the *unprocessed* data; the "Sample = Stage Two" section represents the taxonomy topics that have come into Stage Two after exporting from the top topic of the Stage One; and the third section, "Sample = Topic 8", represents the taxonomy topics found in the top Stage Two topic (Topic 8).

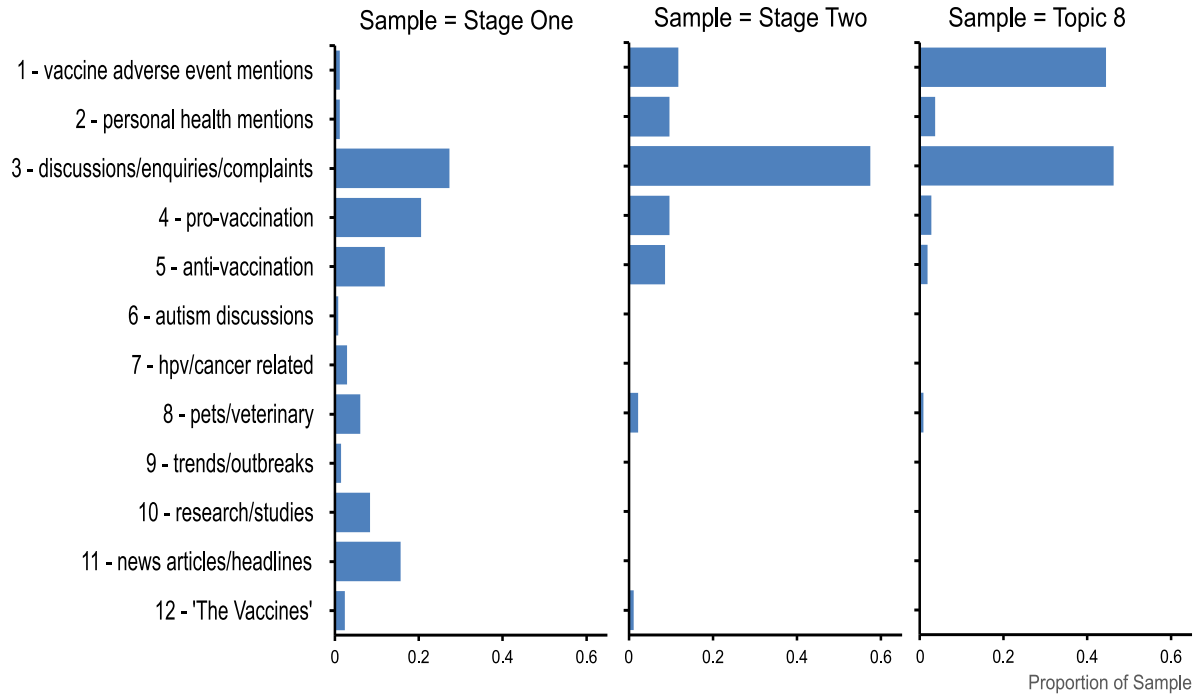


Figure 27: Proportions of Topic per Stage samples

The figure shows that as the topic modelling proceeds the proportion of tweets that contain VAEM (taxonomy topic 1) and similar posts increases. Initially, in the Stage One sample, VAEM is a very small proportion of the data, and topics are distributed over all the data with the highest proportion being in taxonomy topic 3 — discussions, enquiries, and complaints. In the Stage Two sample in the middle of the chart many of the taxonomy topics have been reduced or eliminated, and although topic 3 still dominates there is a far higher proportion of VAEM. The best VAEM topic of Stage Two was Topic 8, which is rendered in the third sample in the chart — now VAEM is proportionally the largest topic and the spread of topics is reduced to mostly VAEM, discussions, and some non-adverse reaction personal health mentions. These observations are discussed in detail in the following charts, which present the same information, with the taxonomy topic proportions depicted as percentages.

Figure 28 is a detailed view of the same Stage One data that is depicted in Figure 27, which is from the DMM 14-topic 1,000-tweet sample. It shows that only 1.2% of the data can be allocated to VAEM. Personal health mentions also occupy 1.2% of the data, while the largest group at 27.3% is topic 3 - discussions, enquires and complaints about vaccinations. Pro and anti-vaccination tweets, and news articles and headlines occupy the next largest groups.

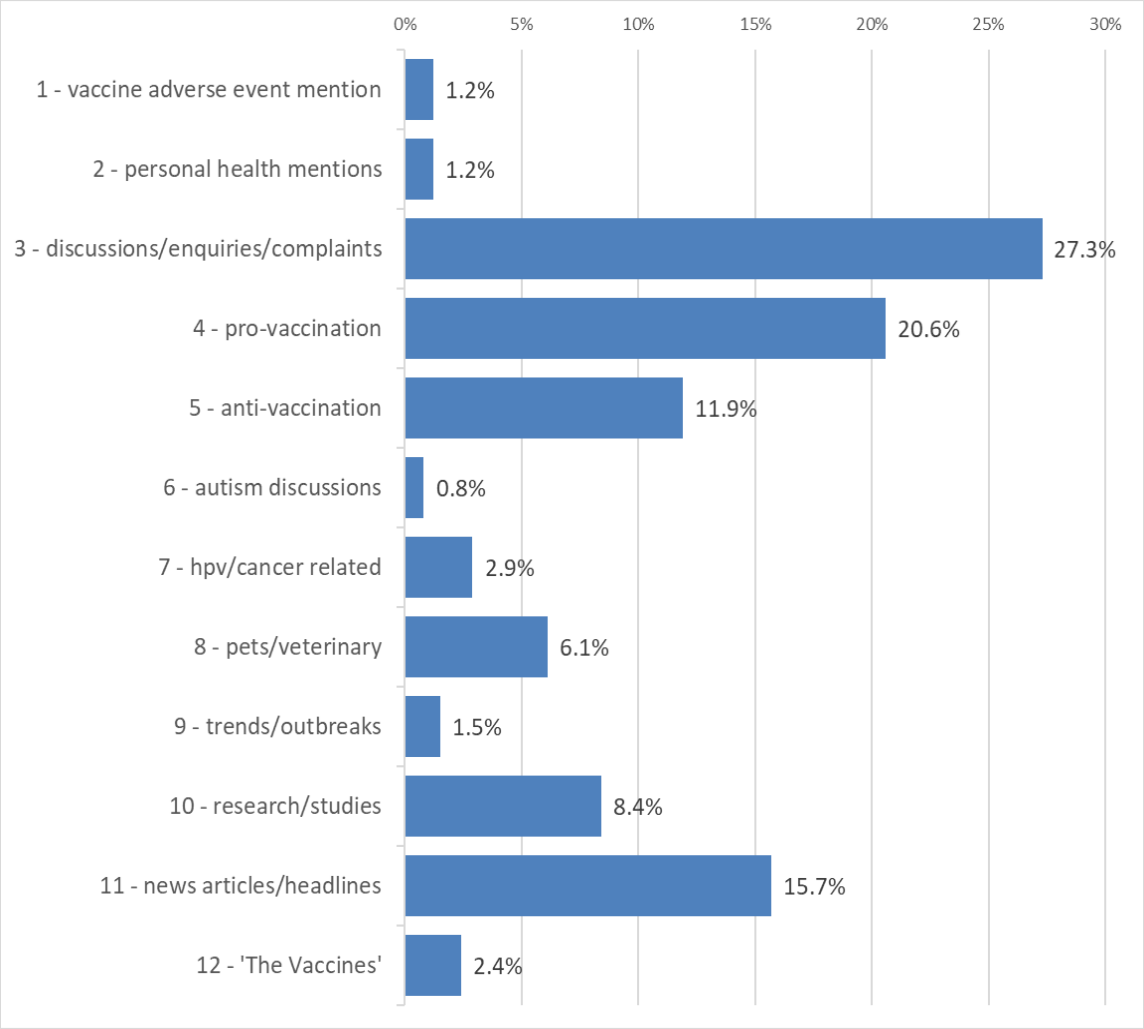


Figure 28: Distribution of taxonomy in 1,000 tweet sample of stage one data

Figure 29 shows the detail of distribution of taxonomy topics in Stage Two sample shown in Figure 27, which used the 1,000-tweet sample from the second stage DMM 9-topic model. To reiterate, this is a sample from *all* of stage two data. Now VAEM occupy 8% of the data, personal health mentions are 4.7% and discussions are 64.2%. The number of other taxonomy topics is much reduced, occupying just 23.1% - the trend is that the data is much more heavily focussed on the types of tweets that might contain VAEM.

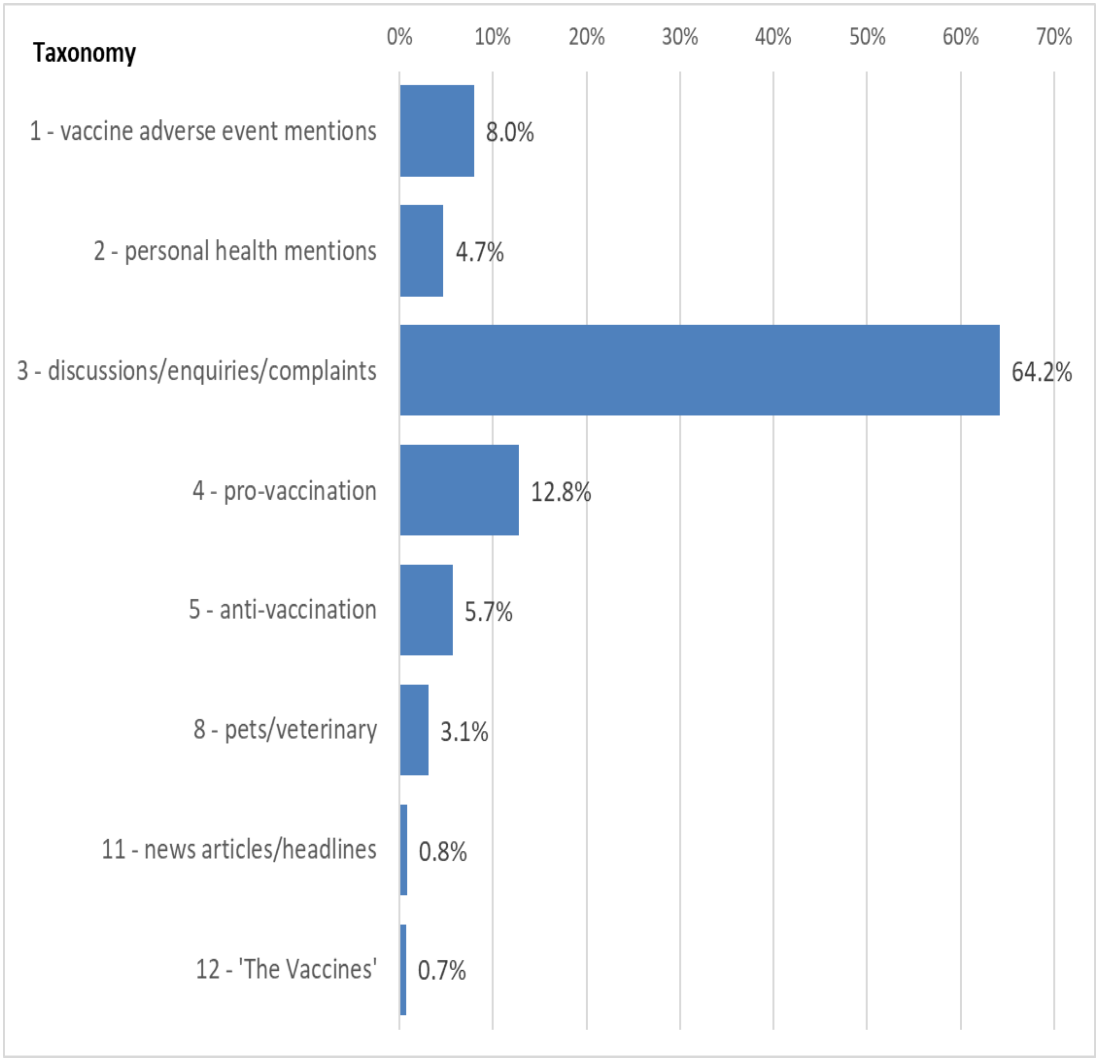


Figure 29: Distribution of taxonomy in 1,000 tweet sample of stage two data

Figure 30 is the detail of the third sample shown in Figure 27, which uses a fresh 1,000-tweet sample just from Topic 8 of the Stage Two data — now VAEM occupy 51.5% of the data, which is more than any other topic. It is quite a lot less than the actual proportion of 75.1% which was counted when the data was fully labelled (see next section), but this *was* only a sample. The tweets are nearly all VAEM, personal health mentions, or discussions. Data from this topic is very amenable to further classification work to identify the VAEM with greater precision.

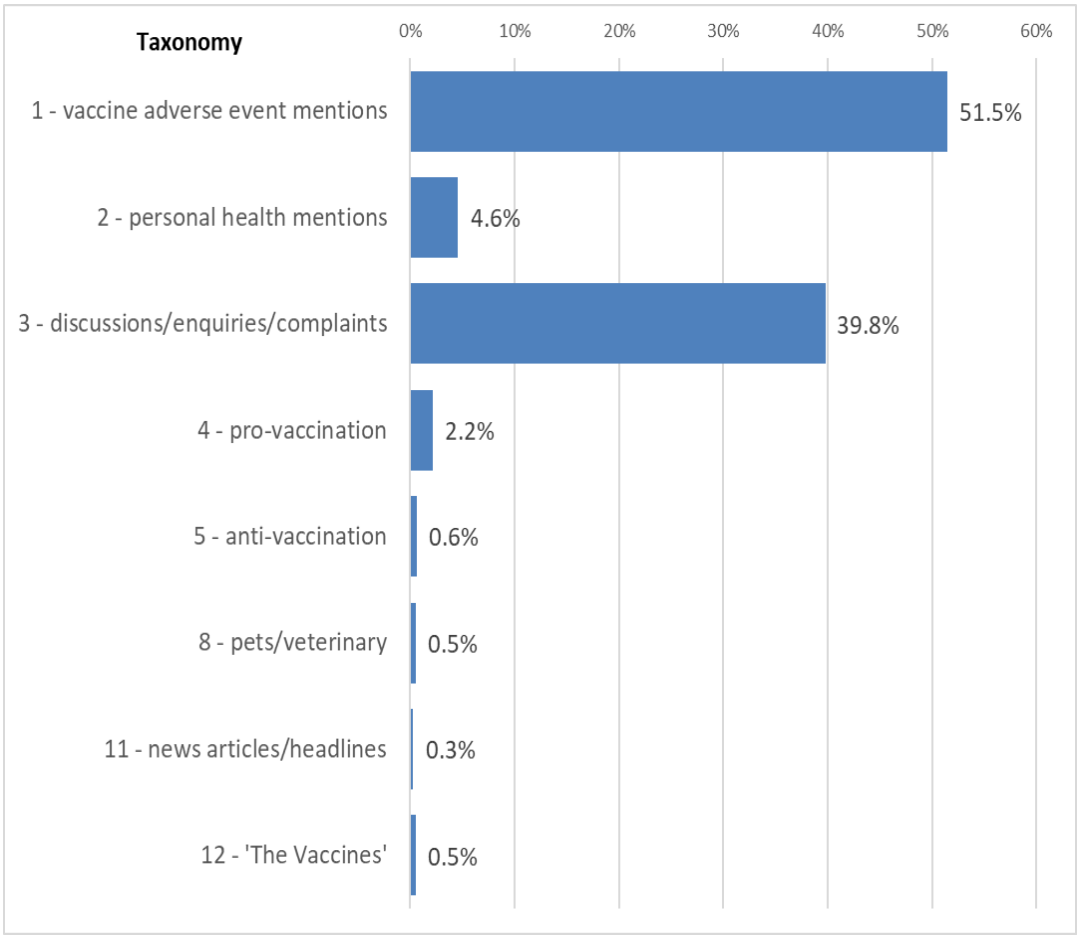


Figure 30: Distribution of taxonomy in 1,000 tweet sample of Topic 8

7.2.2 Verifying effectiveness with label distributions

The previous subsection described how the effectiveness of the two-stage topic modelling approach was verified with samples, this subsection describes how effectiveness was verified by observing the distribution of *labelled* VAEM in the various topics of the topic models. This approach was made possible after processing the second phase data (Section 4.7) through both

stage one and stage two DMM topic models. Because that data was subsequently fully labelled, the distribution of the VAEM in the various stage 2 topics was observable.

That is, the 14-topic DMM model previously ranked as the best model was applied to the 359,535 posts in that dataset, and 80,372 records that the model put into Topic 13 (the best VAEM topic of the model) were retained. The second phase DMM 9-topic model was applied to these, and *all* the records were retained, along with their second-phase topic number.

Figure 31 shows the distribution of these records. It confirms that most of the VAEM containing documents are captured by Topic 8, the VAEM topic. That is, Topic 8 with 6,320 records contains 77.5% of the total of 8,157 VAEM records. It also shows that, compared to other topics, Topic 8 effectively isolates VAEM in proportion to non-VAEM in the topic - the 6,320 VAEM records in the topic are 1.24 times as many as the 5,075 non-VAEM records.

Figure 31: Labels per topic, second stage topic model over the second phase dataset

The numbers also confirm that all other topics have much fewer VAEM documents with the best of them proportionally being Topic 9 with 534 to 6,953 — that is, a proportion of 0.08 VAEM records — significantly less than Topic 8's 1.24 proportion of VAEM. Following that is Topic 7, which although having a greater number of VAEM at 720 also has more non-VAEM

at 12,476 — a proportion of 0.06. Figure 32 presents these relationships in a summary form, which highlights the degree to which Topic 8 contains the majority of VAEM while simultaneously excluding most of the non-VAEM records, which are mostly found in other topics.

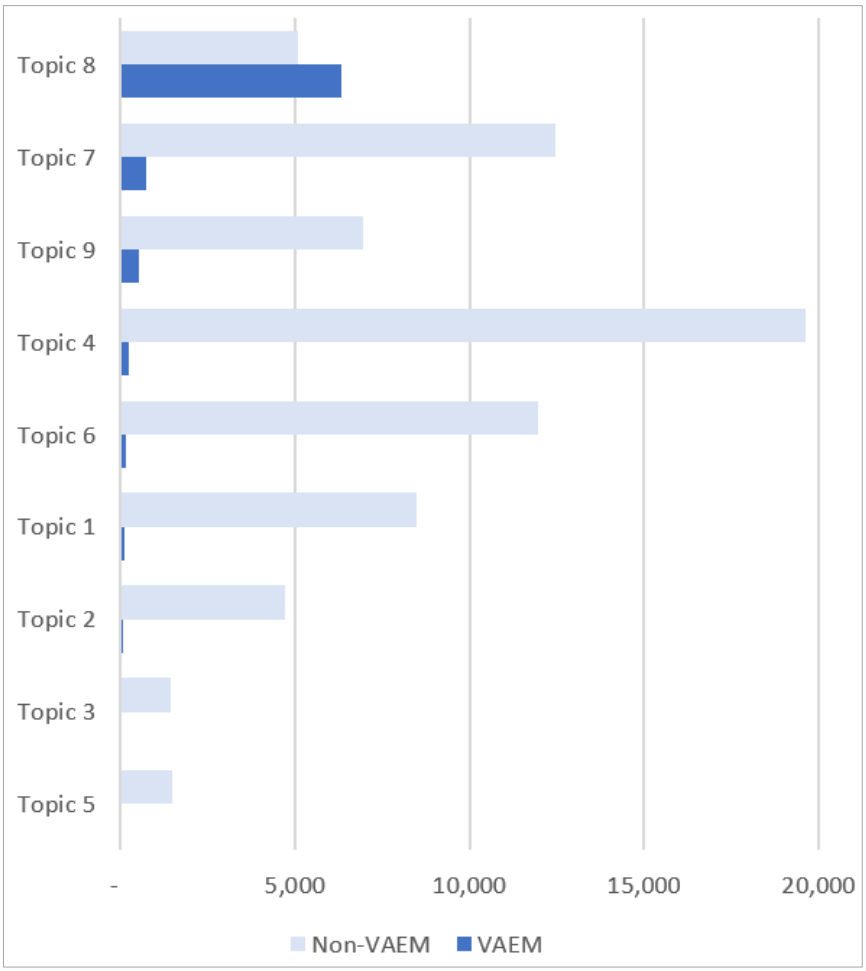
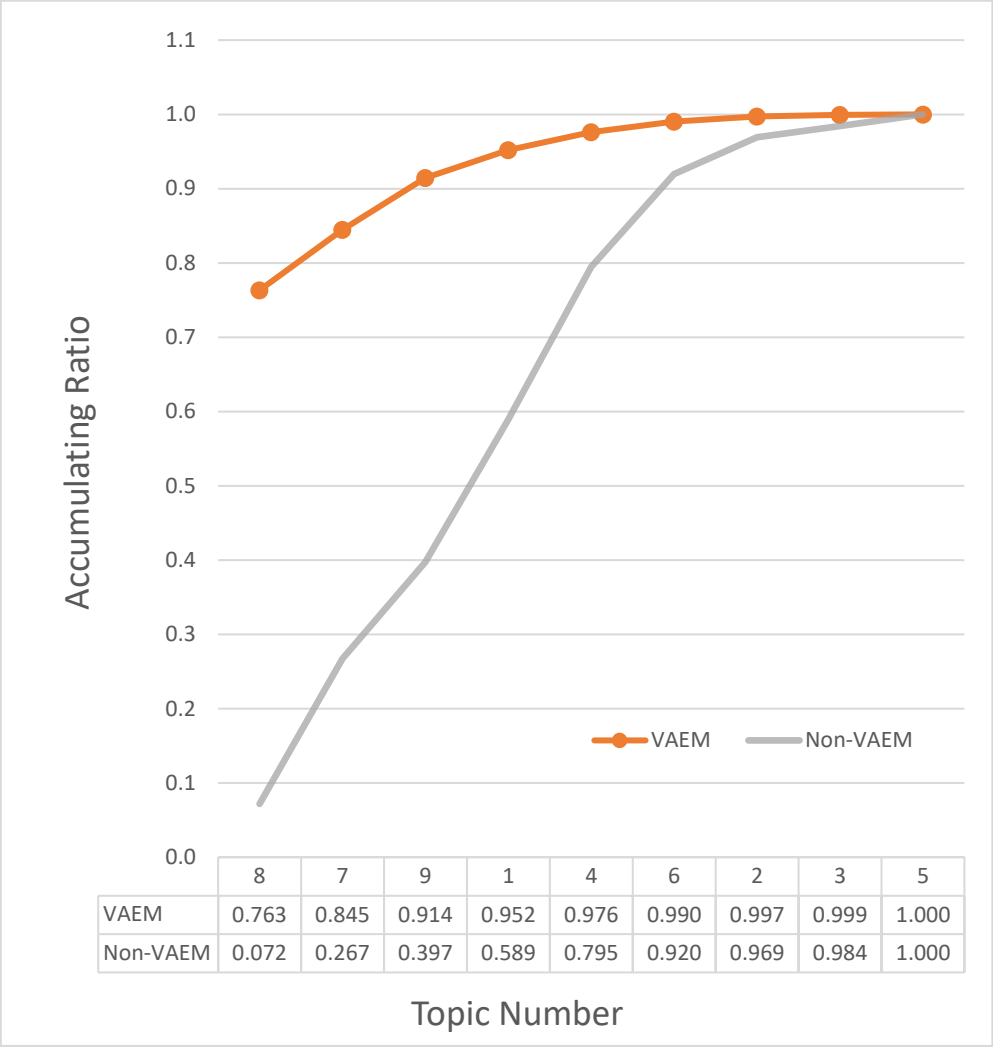


Figure 32: Summary of second stage VAEM distribution in topics

The following charts consider this information as ratios, and over all data — that is, when stage two topic modelling numbers are combined for both phases of data collection.

Figure 33 illustrates the effect of the increasing ratio of labelled VAEM and accompanying labelled non-VAEM as topics are added to the best Topic 8 of the second stage of topic modelling. That is, given that Topic 8 contains the majority of VAEM, the chart depicts the effect of starting with Topic 8, and incrementally adding other topics until the most VAEM are retained. Topics are added in the order of their amount of VAEM to non-VAEM ratio, starting with Topic 8, followed by Topic 7, then 9 etc. Topic 8 contains 0.763 of all labelled VAEM; Topic 7 contains 0.082, which accumulates to 0.845 etc. The lower line is the accumulating ratio of non-VAEM in each topic, as topics are added then the non-VAEM ratio also increases.

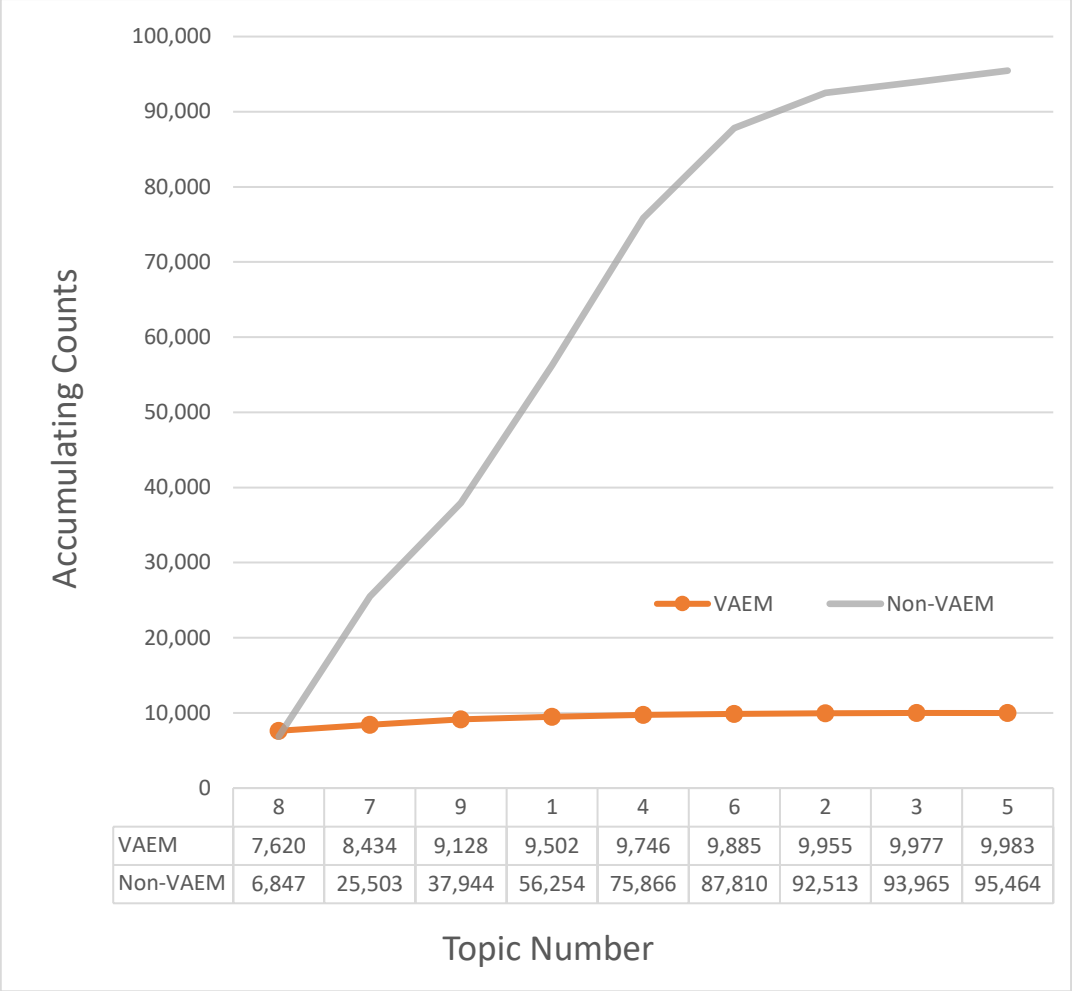
Retaining just Topic 8 would obtain 0.763 of all VAEM and only 0.072 of all non-VAEM; adding Topic 7 would increase to 0.845 of all VAEM and 0.267 of all non-VAEM, etc.



	8	7	9	1	4	6	2	3	5
VAEM	0.763	0.845	0.914	0.952	0.976	0.990	0.997	0.999	1.000
Non-VAEM	0.072	0.267	0.397	0.589	0.795	0.920	0.969	0.984	1.000

Figure 33: Stage Two topics - Accumulating ratio of VAEM vs non-VAEM

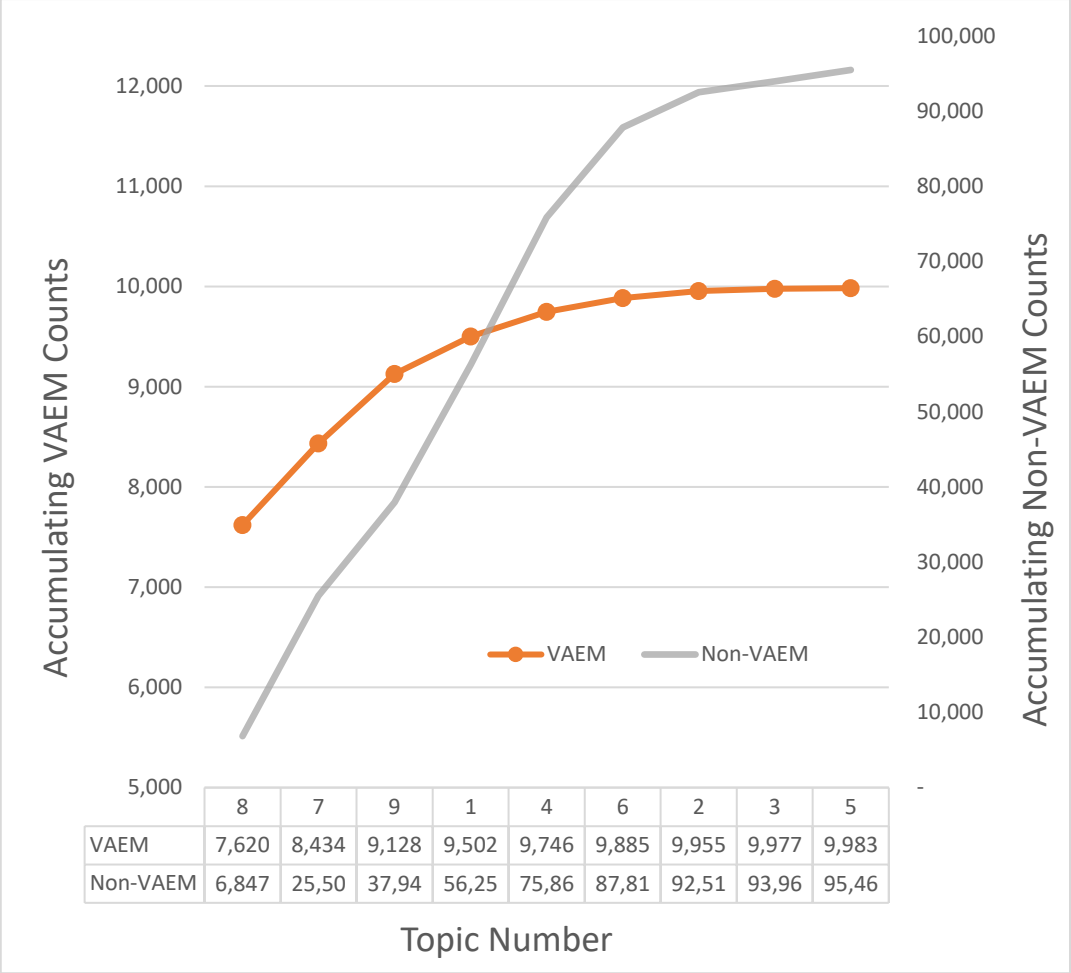
The actual numbers involved are depicted in Figure 34, here it can be seen that the 0.072 of the non-VAEM in Topic 8 represents 6,847 records, which is less than the 7,620 records that 0.763 of VAEM represents; but when Topic 7 is added the non-VAEM records increase by 18,656 to 25,503 records while the VAEM increase by only 814 to 8,434 records. As topics are added the record counts in the non-VAEM groups soon overwhelm those of the VAEM.



	8	7	9	1	4	6	2	3	5
VAEM	7,620	8,434	9,128	9,502	9,746	9,885	9,955	9,977	9,983
Non-VAEM	6,847	25,503	37,944	56,254	75,866	87,810	92,513	93,965	95,464

Figure 34: Stage Two topics — Accumulating counts of VAEM vs non-VAEM

Figure 35 shows this data with separate scales to make it easier to visualise the actual numbers as topics are added, with the caveat that attention must be paid to scale differences.



	8	7	9	1	4	6	2	3	5
VAEM	7,620	8,434	9,128	9,502	9,746	9,885	9,955	9,977	9,983
Non-VAEM	6,847	25,50	37,94	56,25	75,86	87,81	92,51	93,96	95,46

Figure 35: Stage Two topics — Accumulating counts scaled

This section's analysis emphatically demonstrates that the approach of identifying the best performing topic models based on scoring a small sample of labelled data was highly effective. The topic models and their best topics which were identified using that technique performed very well when applied to new data, and the conclusions reached about the best performing topics when using the small, labelled sample are substantiated.

7.2.3 Utilising topic model outputs

Considering the analysis above, using Topic 13 from the first stage of topic modelling is a successful filtering strategy. The second stage of topic modelling also succeeds in further filtering the data, with 75% or more of the potential safety being isolated into Topic 8.

Therefore, a choice must be made when using this approach to either take just the top topic from a first stage of topic modelling output as data and train classifiers to identify the signal of interest; or to utilize second stage topic modelling to further filter the data — looking for a balance that contains the majority of the signals of interest without also containing too many

non-signals. If the latter approach is taken then a decision must be made whether the large percentage of a signal delivered by the best topic in stage two is sufficient, or if other signal-containing topics need to be included even though doing so adds a lot of unwanted data — the summaries of the numbers obtained in this section should assist with making this decision.

For training classifiers in *this* research, both approaches were used: the initial phase of classification training used just data from the second stage top topics 8, 9 and 1 as labelled input data, but the final phase of classification took all the stage one Topic 13 as input and labelled all of it; then added it to the initial labelled data. When training classifiers it was determined that a balanced dataset was preferable, so most of the non-VAEM records were subsequently discarded to get a similar number to the VAEM. Balancing the data was only an option because the data was manually labelled for training — for ongoing purposes data delivered by one or two stages of topic modelling must be used without labelling, then using the trained classifiers to identify the VAEM.

7.3 Evaluating classifiers effectiveness

7.3.1 Comparative charts

Assessments of classifiers were provided in Chapter 6, which presented tables and charts of the scores obtained by increasingly sophisticated classifiers applied to different sizes of input data and discussed a baseline rule-based classification approach. Additionally, Appendices D and E contain evaluations of experiments with using embeddings with traditional classifiers and with utilizing feature engineering approaches. This section brings these various results together into charts to enable a visual understanding of the performance of the classifiers relative to each other and to the available data. The comparative F1-Scores and Adjusted F1-Beta Scores used in the charts and tables are over the Imbalanced Victorian Test dataset, which functioned as a benchmark throughout evaluation.

Training in the initial phase of classification evaluation used the smaller 3,519 record dataset and Figure 36 shows the results from this training phase. The chart arranges the models from left to right in increasing order of their F1-Scores. The rule-based model is included for reference but is not included in subsequent analysis. Also incorporated are the results of an experiment with Word2Vec centroids — represented by the Extra Trees Centroids model — these classifiers were not described in Chapter 6, as the discussion there focussed on standard models.

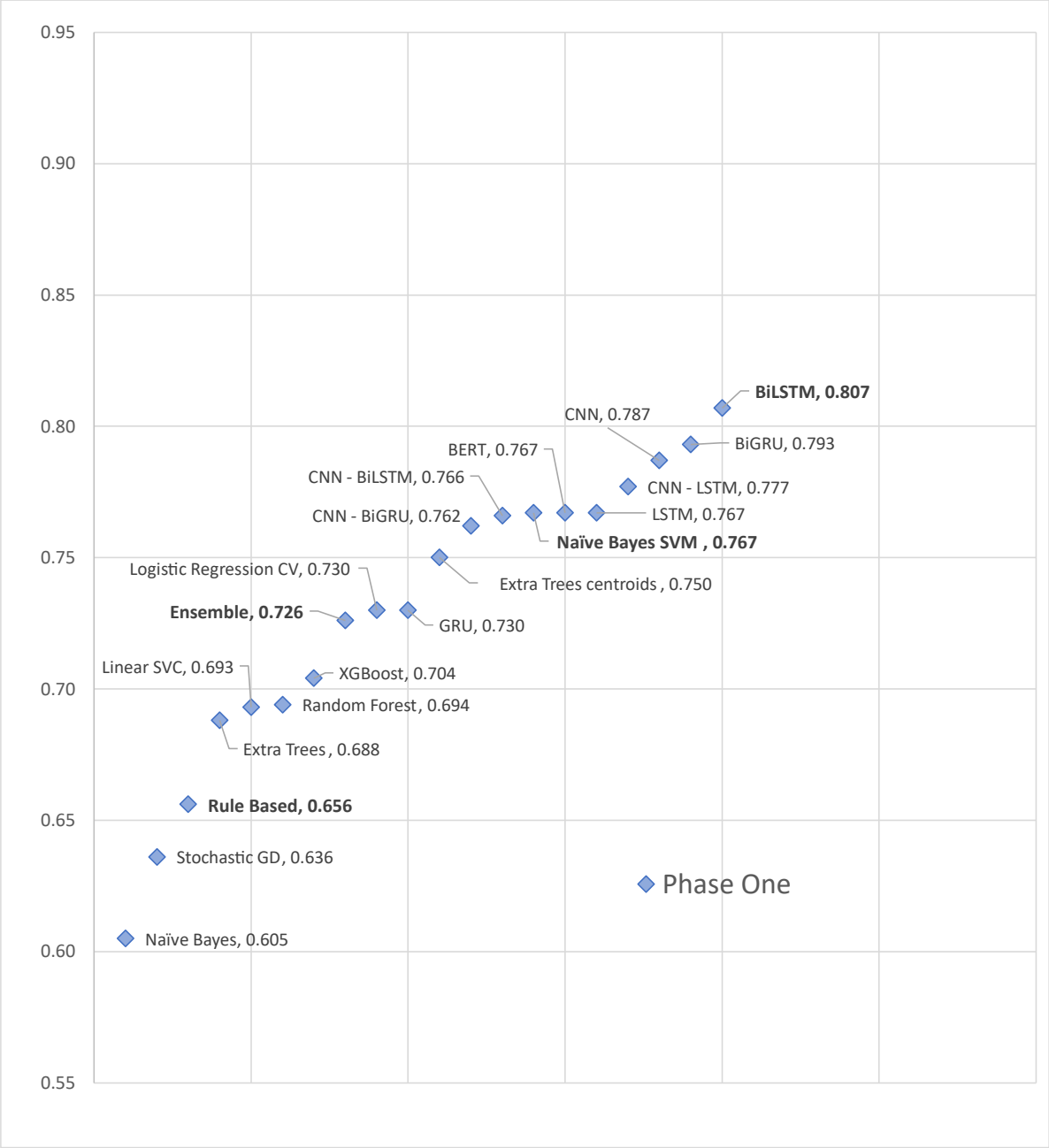


Figure 36: F1-Scores Victorian test data - first classification phase

The deep learning models surpassed the traditional classifiers by scores between 0.06 and 0.11 and are a distinct group in the top portion of the chart. The best of these were neural networks that were trained from scratch - CNNs or CNN hybrids like CNN-BiLSTM and CNN-BiGRU (i.e., a CNN combined with a bi-directional LSTM or GRU). The pure sequence-based classifiers like the LSTM and GRU did not perform as well as their CNN hybrid variants. The fine-tuned transfer-learning approach was trialled using the BERT Transformer, but it did not perform as expected and highlighted the problem of working with a small dataset.

On the other end of the chart, the rule-based classifier is a benchmark with an F1-Score of 0.656. The Naïve Bayes and Stochastic Gradient Descent (SGD) models were both poorer than it, however the poor result of the SGD model was on the Imbalanced test data only, on the Balanced dataset it was one of the best models, and it significantly improved when more training data was made available. The Naïve Bayes SVM model was the best traditional classifier, with an F1-Score of 0.767, but on the Balanced test data it was not a strong performer — these findings are discussed in Chapter 6. The embeddings-based Extra Trees Centroids model with an F1-Score of 0.750 was one of the best performing traditional classifiers, but only in this phase of training - when more training data became available it was no longer prominent. The third-best traditional classifier was the ensemble of standard models at 0.726.

Figure 37 charts the results of re-training all models and the additionally included Transformer models on the 20,077 records of the second phase dataset.

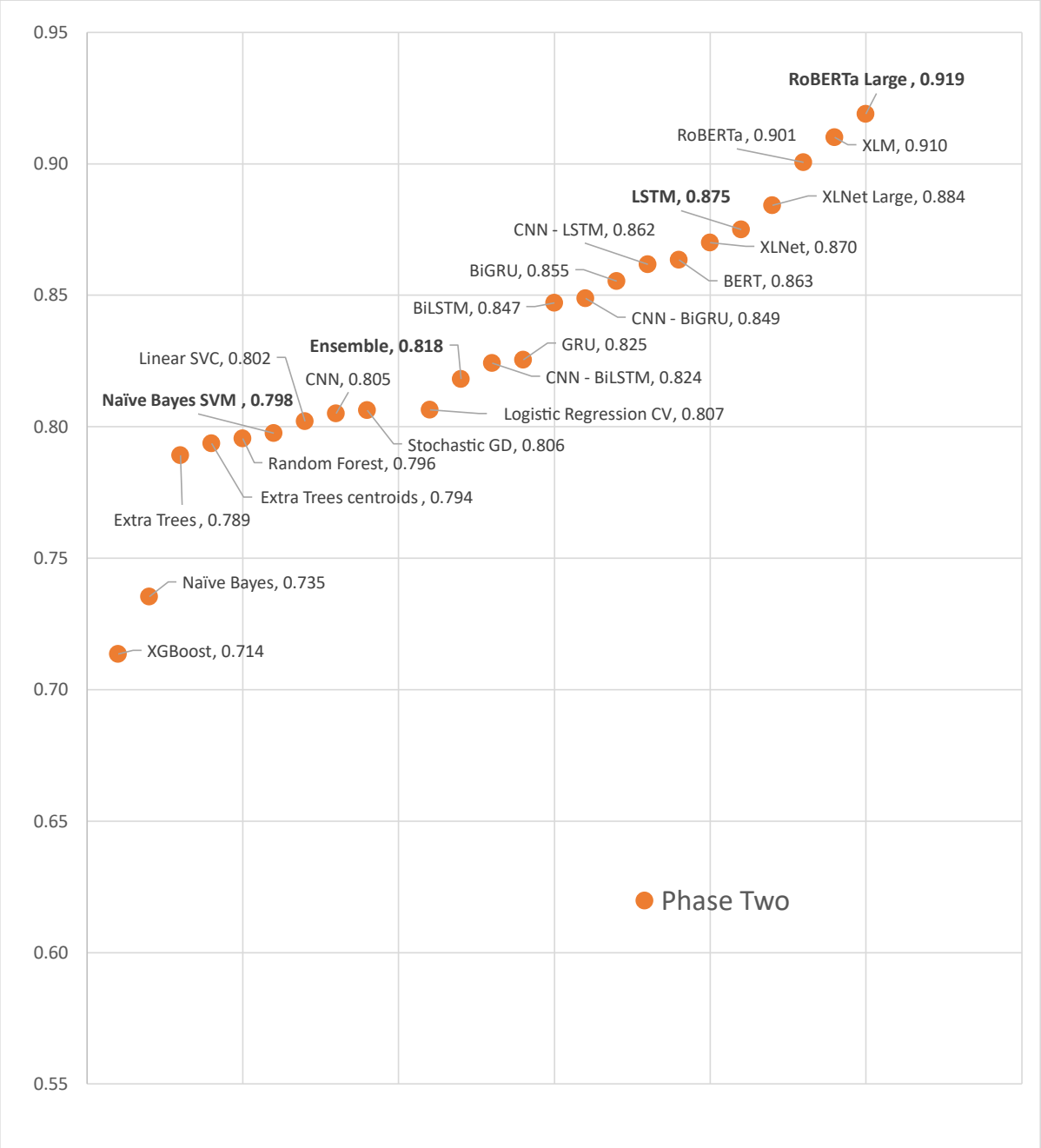


Figure 37: F1-Scores Victorian test data - second classification phase

The chart shares the same scale as Figure 36 to enable an easier comparison of the trends. In Phase Two, with more data to work with, all the classifiers performed better, and there was no longer an abrupt division between the standard and deep learning classifiers. Six of the traditional classifiers outperformed the Extra Trees centroids model which only increased its F1-Score by 0.05, and these six models also outperformed the CNN classifier. Apart from the Extra Trees centroids model in Phase One, the best traditional classifier in both phases was the Ensemble, this was due to its greater precision through the elimination of false positives. SGD and Logistic Regression CV models were the next best traditional classifiers in this phase.

The LSTM model is shown to be the outstanding performer in the neural networks trained from scratch, outperforming both BERT and the XLNet Transformer model. The CNN-LSTM hybrid was the next deep learning model trained from scratch. Beyond these, the rest of the Transformer-based models dominated, with the RoBERTa Large model leading with an F1-Score of 0.919.

Figure 38 combines the data from these two charts into one. It is included to assist with understanding how much each model has improved with the addition of more training data. The models are organized from left to right in order of their second phase scores, which use round markers. Each model's first phase score is located vertically below its higher second phase score and use diamond shaped markers. This means the first phase scores in the chart are no longer arranged in their first phase performance order from left to right, resulting in an irregular pattern to the points. Some models do not have comparative points — the rule-based model and the manually tuned CNN only appear in the Phase One data; and except for BERT the Transformer models only appear in the second phase data. Observing the vertical differences within a single model in each phase assists in understanding how responsive it was to an increase in data.

The chart shows the performant Extra Trees Centroids model from Phase One has a 0.05 increase in performance with the extra data, but that the six traditional classifiers arranged above it have enjoyed increases of 0.11 or more and have overtaken it. The Stochastic Gradient Descent model shows a massive 0.17 increase. Of note among the models are the RoBERTa Large model and the increases of F1-Scores for the Ensemble and the LSTM model — they represent the best models amongst the traditional classifiers and the neural networks trained from scratch.

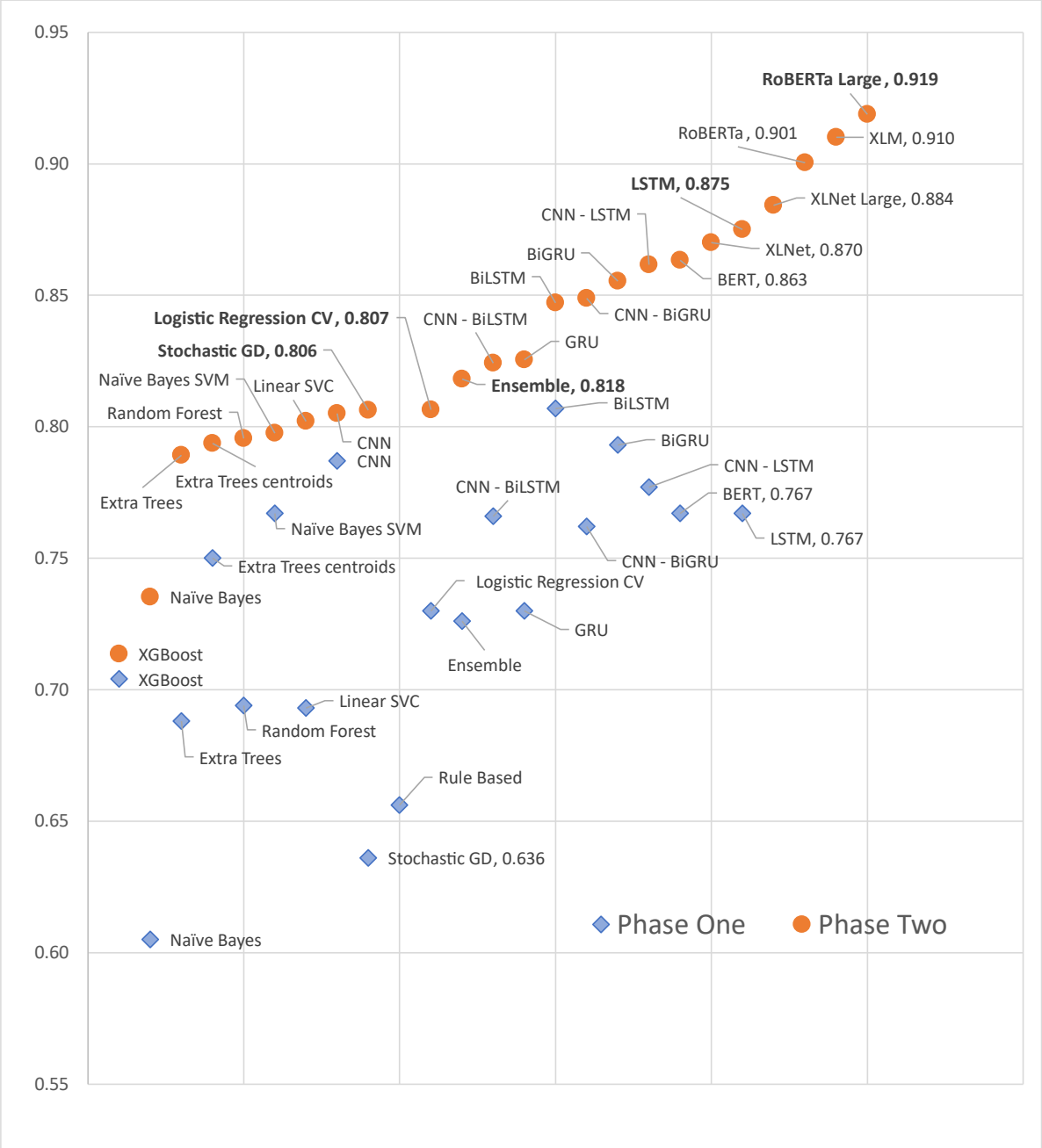


Figure 38: F1-Scores Victorian test data - two classification phases

7.3.2 Detailed analysis of classifier scores

The following tables illustrate these results from Phase One of Classification, in terms of the confusion matrix values of True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), and the accompanying Precision, Recall, F1-Score and Adjusted F-Score (F1-Beta calculated with a beta of 1.3). The figures are evaluated on the Imbalanced Victorian test set - which consists of 614 records, 90 VAEM — the positive label, and 524 non-VAEM — the negative label.

Table 36: Phase One - Imbalanced test dataset - Confusion Matrixes and Scores

Model	TP	TN	FP	FN	Precision	Recall	F1	F1-Beta
BiLSTM	73	506	18	17	0.8022	0.8111	0.8066	**0.8078**
BiGRU	69	509	15	21	0.8214	0.7667	0.7931	0.7861
CNN	72	503	21	18	0.7742	0.8000	0.7869	0.7902
CNN-LSTM	68	507	17	22	0.8000	0.7556	0.7771	0.7715
Naïve Bayes SVM	66	508	16	24	0.8049	0.7333	0.7674	0.7584
BERT	**79**	487	37	11	0.6810	0.8778	0.7670	**0.7927**
LSTM	69	503	21	21	0.7667	0.7667	0.7667	0.7667
CNN-BiLSTM	72	498	26	18	0.7347	0.8000	0.7660	0.7744
CNN-BiGRU	77	489	35	13	0.6875	0.8556	0.7624	0.7843
Extra Trees centroids	72	494	30	18	0.7059	0.8000	0.7500	0.7622
GRU	77	480	44	13	0.6364	0.8556	0.7299	0.7584
Logistic Regression CV	77	480	44	13	0.6364	0.8556	0.7299	0.7584
Ensemble	77	479	45	13	0.6311	0.8556	0.7264	0.7557
XG Boost	75	476	48	15	0.6098	0.8333	0.7042	0.7334
Random Forest	77	469	55	13	0.5833	0.8556	0.6937	0.7291
Linear SVC	79	465	59	11	0.5725	0.8778	0.6930	0.7325
Extra Trees	76	469	55	14	0.5802	0.8444	0.6878	0.7221
Rule-Based	60	496	28	30	0.6818	0.6667	0.6742	0.6722
Stochastic GD	77	449	75	13	0.5066	0.8556	0.6364	0.6811
Naïve Bayes	**82**	425	99	8	0.4530	0.9111	0.6052	0.6622

Table 36 above displays these figures from Phase One of the Classification evaluation, in descending order of F1-Score. The F1-Beta (beta of 1.3) generally follows same trend as the F1-Score, except for models such as the BERT model where a much higher recall results in a higher F1-Beta. Compare the BERT and Naïve Bayes SVM models: the NBSVM has a higher F1-Score but markedly lower F1-Beta, because its recall is much lower, with its high precision being at the expense of recall. Compared to the top-scoring BiLSTM, the BERT model again has a higher recall, but the other model is better because of its balance of recall and precision — the BiLSTM model's recall is somewhat lower than many of the models below it, but its precision is higher than most.

The top scoring models are clearly better at detecting language differences between VAEM and not, which is reflected in a reduction of false positives in these models, leading to higher precision.

Table 37 looks at the equivalent information from the second phase of Classification, with the exception that rule-based model did not feature in this phase and the table includes additional Transformer models.

Table 37: Phase Two - Imbalanced test dataset - Confusion Matrixes and Scores

Model	TP	TN	FP	FN	Precision	Recall	F1	F1-Beta
RoBERTa Large	85	514	10	5	0.8947	0.9444	0.9189	0.9253
XLM	86	511	13	4	0.8687	0.9556	0.9101	0.9213
RoBERTa	86	509	15	4	0.8515	0.9556	0.9005	0.9140
XLNet Large	84	508	16	6	0.8400	0.9333	0.8842	0.8963
LSTM	77	515	9	13	0.8953	0.8556	0.8750	0.8699
XLNet	87	501	23	3	0.7909	0.9667	0.8700	0.8929
BERT	79	510	14	11	0.8495	0.8778	0.8634	0.8670
CNN-LSTM	81	507	17	9	0.8265	0.9000	0.8617	0.8712
BiGRU	71	519	5	19	0.9342	0.7889	0.8554	0.8373
CNN-BiGRU	73	515	9	17	0.8902	0.8111	0.8488	0.8388
BiLSTM	72	516	8	18	0.9000	0.8000	0.8471	0.8345
GRU	78	503	21	12	0.7879	0.8667	0.8254	0.8356
CNN-BiLSTM	68	517	7	22	0.9067	0.7556	0.8242	0.8055
Ensemble	72	510	14	18	0.8372	0.8000	0.8182	0.8134
Logistic Regression CV	75	503	21	15	0.7813	0.8333	0.8065	0.8132
Stochastic GD	77	500	24	13	0.7624	0.8556	0.8063	0.8184
CNN	72	507	17	18	0.8090	0.8000	0.8045	0.8033
Linear SVC	75	502	22	15	0.7732	0.8333	0.8021	0.8099
Naïve Bayes SVM	65	516	8	25	0.8904	0.7222	0.7975	0.7768
Random Forest	72	505	19	18	0.7912	0.8000	0.7956	0.7967
Extra Trees Centroids	75	500	24	15	0.7576	0.8333	0.7937	0.8035
Extra Trees	73	502	22	17	0.7684	0.8111	0.7892	0.7947
Naïve Bayes	75	485	39	15	0.6579	0.8333	0.7353	0.7582
XG Boost	61	504	20	29	0.7531	0.6778	0.7135	0.7039

With more data to train on, the second phase results are all much better, due to reductions in both false positives and false negatives. For instance, the LSTM model has reduced false positives from 21 to 9, and false negatives from 21 to 13, and has overtaken the CNN hybrid models and even outperforms the BERT and the XLNet Transformer models. The Naïve Bayes SVM model however, while halving its false positives from 16 to 8, has increased its false negatives from 24 to 25 — so it has gained precision at the expense of recall, and overall has not improved as much as other models which now surpass it. The experimental Extra Trees Centroids which did reasonably well on the Phase One training data is now towards the bottom of the table.

Table 38 lists all the Classifiers assessed in Phase Two of classification with their combined test scores. That is, the predictions from the larger Balanced Test set introduced in Phase Two added to the predictions from the Imbalanced Victorian test set from Phase Two, which were explored above. For some classifiers, the figures are lower than those achieved purely on the

Imbalanced test set, but these figures should be a truer indicator of the classifiers' capabilities. Note that the LSTM model is no longer the best performer from the neural nets trained from scratch, the bi-directional GRU is now the best performer of those models. RoBERTa Large is still clearly the strongest model.

Table 38: Phase Two - Combined test datasets - Confusion Matrixes and Scores

Model	TP	TN	FP	FN	Precision	Recall	F1	F1-Beta
RoBERTa Large	494	850	71	27	0.8743	0.9482	0.9098	0.9193
RoBERTa	496	841	80	25	0.8611	0.9520	0.9043	0.9161
XLNet Large	485	852	69	36	0.8755	0.9309	0.9023	0.9095
XLM	475	858	63	46	0.8829	0.9117	0.8971	0.9008
XLNet	478	854	67	43	0.8771	0.9175	0.8968	0.9020
BiGRU	464	863	58	57	0.8889	0.8906	0.8897	0.8900
BERT	476	845	76	45	0.8623	0.9136	0.8872	0.8939
CNN-BiGRU	483	832	89	38	0.8444	0.9271	0.8838	0.8945
BiLSTM	469	843	78	52	0.8574	0.9002	0.8783	0.8838
LSTM	479	830	91	42	0.8404	0.9194	0.8781	0.8883
CNN-LSTM	479	824	97	42	0.8316	0.9194	0.8733	0.8847
CNN-BiLSTM	468	835	86	53	0.8448	0.8983	0.8707	0.8776
GRU	482	813	108	39	0.8169	0.9251	0.8677	0.8817
Ensemble	450	851	70	71	0.8654	0.8637	0.8646	0.8643
Stochastic GD	460	834	87	61	0.8410	0.8829	0.8614	0.8668
Logistic Regression CV	452	844	77	69	0.8544	0.8676	0.8610	0.8626
Linear SVC	456	834	87	65	0.8398	0.8752	0.8571	0.8617
CNN	457	831	90	64	0.8355	0.8772	0.8558	0.8612
Random Forest	445	843	78	76	0.8509	0.8541	0.8525	0.8529
Extra Trees	444	840	81	77	0.8457	0.8522	0.8489	0.8498
Extra Trees Centroids	445	823	98	76	0.8195	0.8541	0.8365	0.8409
Naïve Bayes SVM	412	863	58	109	0.8766	0.7908	0.8315	0.8207
XG Boost	422	848	73	99	0.8525	0.8100	0.8307	0.8253
Naïve Bayes	456	789	132	65	0.7755	0.8752	0.8224	0.8353

7.3.3 Classifier effectiveness

In summary, with the smaller amount of data in Phase One of the classifier testing the best results when measured by F1-Score came with the capacity for higher precision, and hybrids of CNN and bi-directional LSTM or GRU were amongst the best performers here — though the BERT Transformer was better than most if F1-Beta Score was considered. With the addition of more data in Phase Two all models benefitted, the CNN based models continued to perform well but so also did pure LSTM and GRU recurrent neural networks, but the standout

performers were the larger Transformer models — particularly the RoBERTa Large model — its F1-Score was consistently over 0.9 on any of the tests.

7.4 Evaluating effectiveness of the method

This section summarizes the overall effectiveness of the research approach by re-stating previous information in terms of the quantities of tweets that were progressively filtered and identified as having vaccine adverse event mentions. The numbers presented are the combined numbers of data collected and processed over the two data phases of data processing. When presented, the constituent first and second phase numbers are listed within brackets.

Table 39: Summary Topic Modelling counts

Tweets Collected	**811,010**				
- Cleaned	- 122,653				
- Discarded - Stage One	- 570,383				
Remaining	**117,974**	**14.5%**	**of total**		
- Discarded - Stage Two	- 19,083				
Retained and labelled	**98,891**	*Non-VAEM*	*88,900*		
		VAEM	**9,991**	**10.1%**	**of retained data**
		VAEM in other Stage Two topics	*2,367*		
		VAEM in best Stage Two topic	**7,624**	**76.3%**	**of retained VAEM**

Distributions to the completion of the *first* stage of topic modelling:

- 811,010 (400,097 + 410,913) tweets were collected
- 122,653 were cleaned before topic modelling, leaving 688,357
- 570,383 were discarded by stage one of topic modelling, leaving 117,974

Distributions to the completion of the *second* stage of topic modelling:

- 117,974 tweets were processed, which is the total data coming from stage one of topic modelling for both phases of data collection
- 19,083 were discarded — which was those from the first phase of data collection that were not in the top 3 topics of the second-stage topic model
- 98,891 (18,519 + 80,372) were retained and labelled. From the first phase of data collection: the top 3 topics of the second-stage topic model. From the second phase of data collection: all of data filtered by the first-stage topic model

Proportion of VAEM in the 98,891 labelled tweets:

- 88,900 did not contain VAEM
- 9,991 of them contained VAEM
- The proportion of VAEM was 10.1%

Of the 9,991 retained VAEM-containing tweets:

- 7,624 were found to be in the single best second-stage topic
- 2,367 were distributed amongst the remaining 8 topics
- 76.3% of the VAEM were in the best topic

An estimate of the classifier's performance:

- Based on the RoBERTa Large classifier, which accurately identities 90% of the test data, and applying that proportion to all the extracted stage one data:
- 8,992 of the 9,991 filtered first stage VAEM records would be correctly classified

Finally, measuring the combined effectiveness of topic modelling and classification:

- 8,992 VAEM are identified from the original 811,010 records, being 90% of all likely VAEM — with a very high confidence
- Therefore 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.

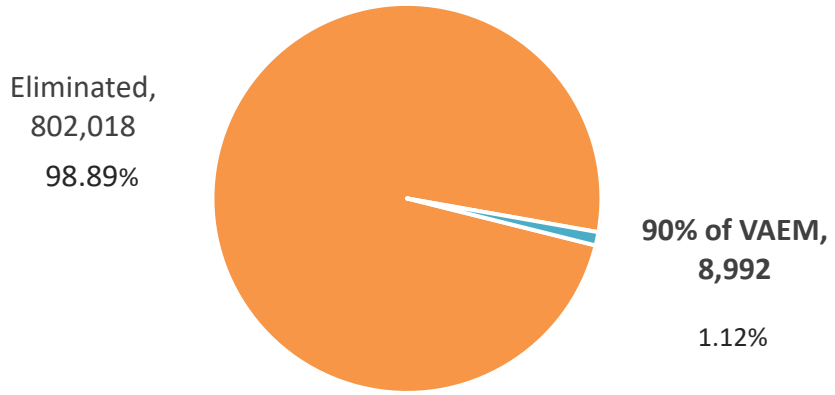


Figure 39: VAEM-Mine method - Capturing of 90% of VAEM

This section presented information about the effectiveness of two stages of topic modelling followed by classification to effectively identify and isolate vaccine adverse event mentions from almost all other vaccine-related Twitter posts. It analysed the topic modelling and classification in detail, then re-stated the information as a sequence of progressive data filtration and concluded that the process followed in this study can reliably extract 90% of vaccine adverse event mentions.

7.5 Word importance analysis

Word differences and similarities of words of the Victorian dataset were visualized, by plotting the relative importance of words in horizontal bar plots, and with word clouds. Firstly, Figure 40 shows the relative importance of the first 30 words, where words are ranked based on a combined scoring between VAEM and non-VAEM texts (which are adjusted to account for the greater volume of texts):

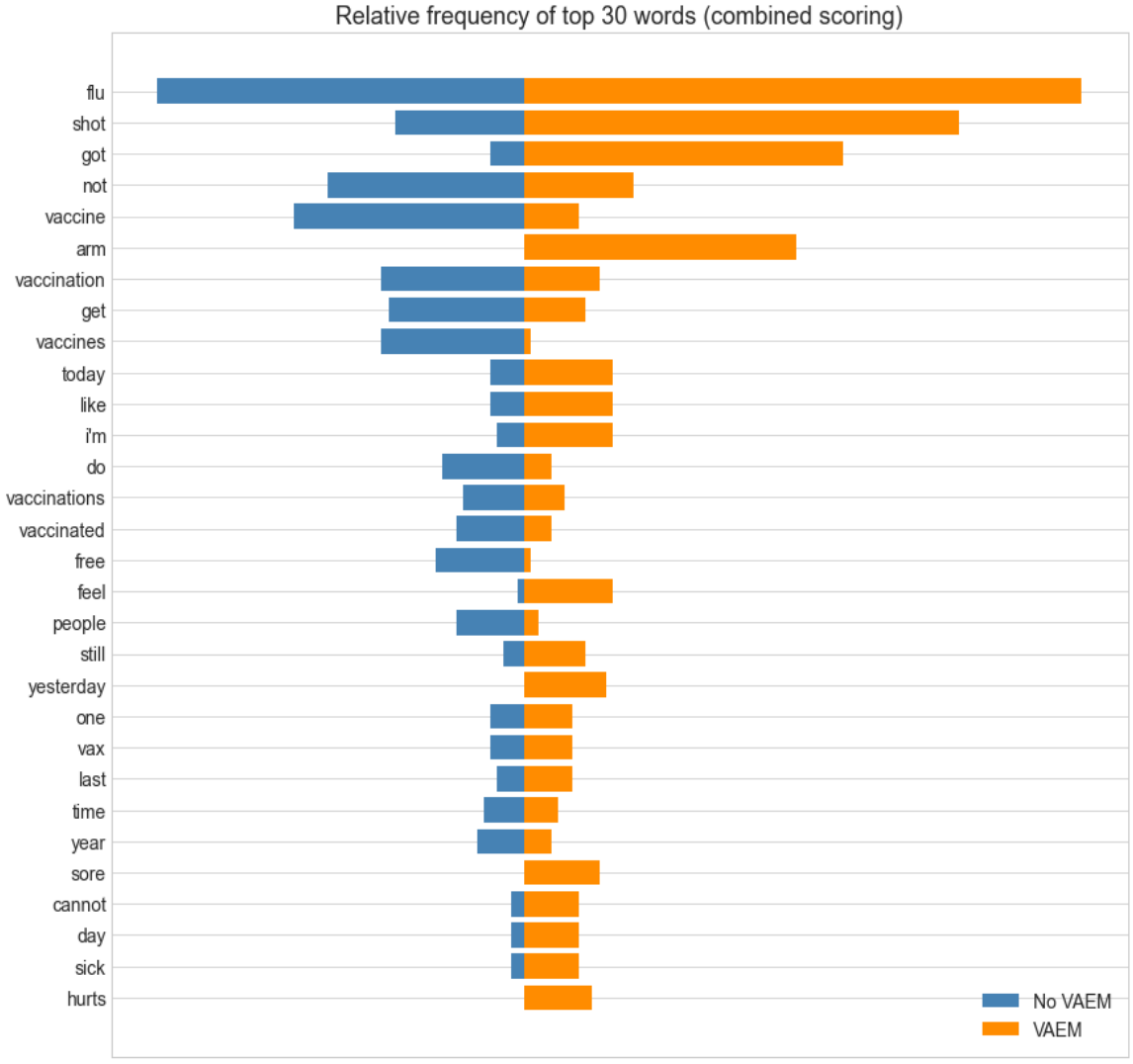


Figure 40: Relative frequency of top 30 words combined

Some words are almost relatively as important, such as ‘flu’ and ‘time’, but there is a greater emphasis on personal meanings of words in the VAEM texts — for instance, ”got” and “yesterday” are likely to relate to something that a person experienced (such as “got a flu shot yesterday”), whereas “get” and “free” are likely to accompany an command to action (such as “get your free flu shot”). “I’m”, “feel” and “sick” (representing “I’m sick” and “feel sick” as potential pairings) are much more strongly emphasized as important in the VAEM set. In fact, words that relate to painful physical symptoms are often exclusively represented in the VAEM set — for instance “arm”, “sore” and “hurts”. Conversely, words that would be used in a general discussion around vaccines have greater emphasis in the non-VAEM set: “vaccine” and “vaccines”, “vaccinations” and “vaccinated”.

This becomes clearer in Figure 41 if the set is taken just from the top 30 words that are important to the vaccine adverse event mentions related to their importance on the non-VAEM side: “lie”, “bed”, “hurting”, “baby” and “right” (as in “right arm”) now also appear in the list, and mostly represented only from the VAEM side.

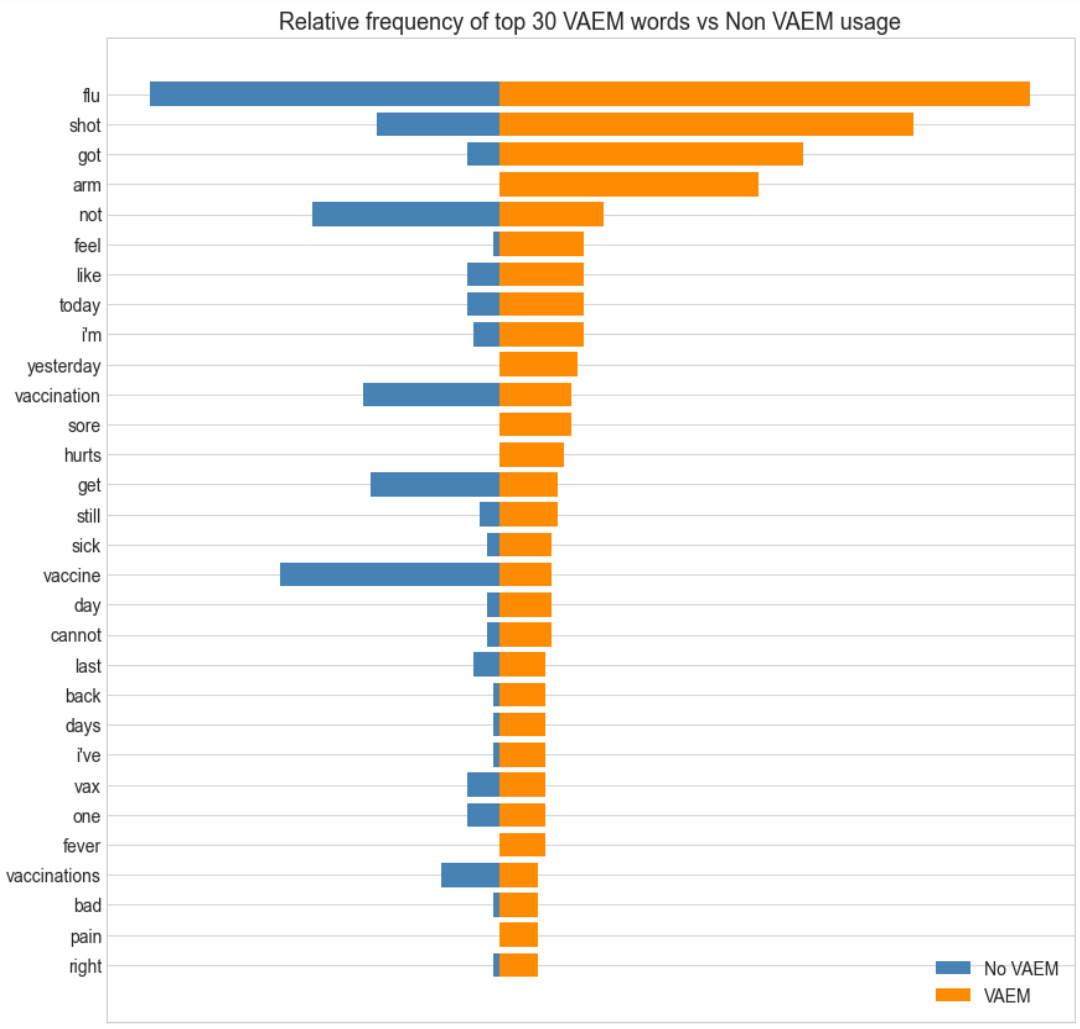


Figure 41: Relative frequency of top 30 VAEM words

Word importance can also be portrayed as word clouds, although they are not quantifiable, they offer a visual interpretation of the relative importance of words. Figure 42 is the word cloud of VAEM, the most significant words are “flu”, “shot”, “got” and “arm”. Other important words are “today”, “yesterday”, “feel” and “sore”, with “sick”, “hurting”, “pain” and “reaction” also gaining visibility. These suggest reporting a recent flu vaccination that has led to a painful arm and feeling sick.

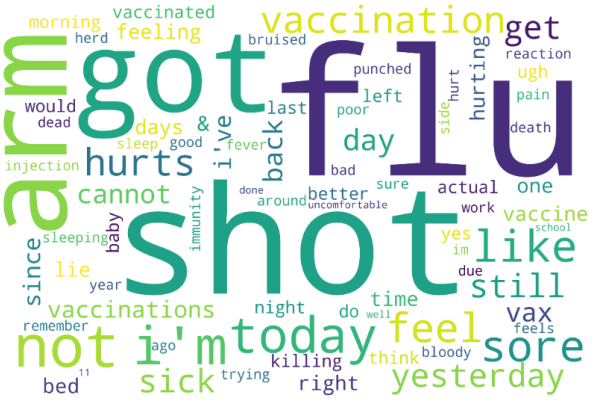


Figure 42: Word cloud of VAEM tweets

The most significant “flu” and “shot” words from the VAEM are also significant in the word cloud of the non-VAEM related tweets - Figure 43. Many of the words are similar, and this illustrates the juxtaposition of the data — there is not much that separates VAEM from non-VAEM after most of the irrelevant data has been eliminated. However, the strongly significant word “got” from the VAEM word cloud has relatively less significance here, instead the equivalent word “get” is strong. This, along with the words “do”, “not”, “people”, “free” and the strength of the words “vaccine” and “vaccination” give the impression that these tweets can be characterized as enjoining people to action, either for or against getting a flu vaccination, or are otherwise discussing vaccines, but are not describing the effects of a recent vaccination, as the VAEM words do.

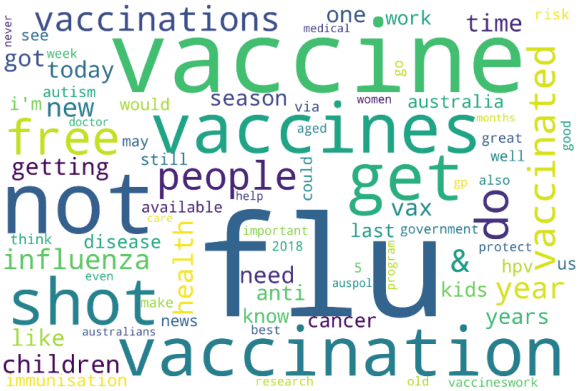


Figure 43: Word cloud of non-VAEM tweets

7.6 Chapter 7 summary

This chapter focussed on evaluating the effectiveness of the research approach for identifying VAEM from the enormous quantity of other vaccine-related tweets. Results were presented using charts and tables that summarized the two-stage topic modelling and classification approaches used. Topic model effectiveness was described in terms of the ability of the models to bring VAEM into one, or a few, topics while not including too much non-VAEM — so that the topics might be used as a filtering mechanism. Analysis of the classifiers compared F1-Score in the context of two phases of data collection and highlighted how classifier effectiveness is linked to data quantity - and that given enough data, the best models can score at 0.80 or more. The inclusion of results from the range of the experiments that were conducted during the research gives context and certainty to the conclusions that were made.

The chapter also examined the nature of the language that differentiates VAEM, to demonstrate that VAEM are semantically distinct, which is the reason that these techniques can be so effective. Appendix G expands on this by looking at the types of VAEM texts that are found in the topics of the models, and it contains textual analyses of the classification errors. The numbers presented in the chapter demonstrate that the research approach is highly effective for identifying VAEM in tweets.

8 Discussion and Conclusion

This research is motivated by the great importance of ensuring that vaccines are safe and effective, so that we can all have confidence in them to help protect us and our loved ones from age-old and emerging deadly diseases. To this end, the research aimed to establish whether social media surveillance can assist with detection of vaccine safety signals, by investigating effective techniques for identifying posts containing Vaccine Adverse Event Mentions (VAEM). The research was grounded in the existing knowledge base of social media mining in public health surveillance. Natural language processing and machine learning techniques used in this domain, especially those used for personal health mention detection, were thoroughly studied, evaluated, and adapted to the research requirement of effectively identifying vaccine adverse event mention posts. It was found that tweets about vaccine reactions are mostly mild personal health mentions, usually accompanied by an informal reference to an influenza vaccine or other common vaccinations. The tweets differed from other related areas like drug reactions - which tended to be more specific about the medication reacted to and with more varied reactions.

The major research contribution was to explain what techniques worked well to elicit these posts, adding to the knowledge of how to go about extracting this information from social media. A pragmatic approach to evaluating topic models was to use F1-scoring on a small number of labelled posts to identify the models that were most effective in placing vaccine adverse event mentions into one topic. Applying a second stage of topic modelling to that original topic helped to further filter out VAEM from the other vaccine-related tweets, so that the task of labelling for classification was manageable. A range of classifiers were assessed, to provide comparative effectiveness measures and to understand that available data quantity must be accounted for when choosing a classifier.

In conclusion, the resulting topic modelling scoring approach was highly effective and was able to be utilized as a core component in the construction of an end-to-end method for extracting VAEM from Twitter posts, but which is also applicable to similar problems. This fulfilled the research expectation that an appropriate method would emerge to not only answer this research need, but also for use in downstream research and solutions.

To evaluate the performance of machine learning algorithms, the research used the standard evaluation metrics of precision, recall and F-measure, including the Adjusted F1-Score. Throughout the development of the classifiers the final arbiter of success was via the F1-Score on an out-of-sample dataset consisting of tweets referencing Victoria, Australia for the period

of 7th February to 7th June 2018 — this was used as a representative of vaccine-related social media conversations in the physical research environment. In summary, models were judged, tuned, and improved using F1-Score, and recall was favoured when deciding between models.

When evaluating classifiers, a range of traditional classifiers and a rule-based method were assessed before moving on to evaluating neural networks. This provided a benchmark and allowed an informed understanding of how much more effective neural networks could be. Additionally, classifiers were evaluated in two phases, firstly training with a smaller dataset, then one which was five times larger. This enabled assessment of the critical effect of dataset size, and that some classifiers did better than others on the smaller dataset, but with more data some of those others were then revealed as outstandingly effective. Specifically, it was found that CNN and CNN-hybrid neural networks did best when dealing with the smaller dataset, but that LSTM and Transformer models were very much better than anything else when trained on the larger dataset.

The research findings are presented as observations and evaluations, reports, and data. The research resulted in a dataset of 692,748 vaccine-related tweets and from these, a binary-labelled dataset of 84,889 vaccine-related tweets, and a subset of this as a balanced dataset of 20,691 tweets. Additionally, the research describes a Twitter-based vaccine-related taxonomy.

8.1 The Research Questions

8.1.1 Aim of the research

How can social media surveillance assist with detection of vaccine safety signals?

The research finds that combinations of readily available NLP tools can be used to extract almost all vaccine adverse event mentions from Twitter data, while eliminating irrelevant posts. The extracted VAEM data has been confirmed as fit for the purpose of safety signal detection by an expert's analysis, and by a comparative trend study. Although the extracted posts were tweets and the techniques used were specifically designed to obtain Twitter posts, the techniques can be adapted to other social media platforms. The research confirms therefore, that social media can assist with detection of vaccine safety signals and can become 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, thus helping to address the deficiencies of timeliness and under-reporting of passive reporting systems.

8.1.2 Research Question 1

What effective techniques can be utilized for identifying posts containing Vaccine Adverse Event Mentions (VAEM)?

It was found that a two-stage topic modelling approach was able to identify up to 99% of Twitter posts containing VAEM, and that classification techniques were able to further isolate them with an over 90% accuracy. A F1-scoring system using a small number of pre-labelled posts was highly effective in identifying the best topic models. Topic modelling as an appropriate filtering technique was aided by the distinctive features of the VAEM Twitter posts, the brevity of tweets, and the use of topic models which do well on one-subject documents. State-of-the-art Transformer deep learning models completed the process of extracting VAEM posts from the texts identified by topic modelling. The techniques and models are applicable to a future software artefact that can be incorporated into a data collection process.

8.1.3 Research Question 2

How can a comprehensive data set be assembled and labelled which will enable both this research and further research into vaccine discourse in social media?

By applying a targeted scoring approach with topic modelling, the VAEM containing tweets were filtered out of the incoming data into a manageable dataset for labelling. A balanced dataset was created containing 20,077 records labelled for the mention of a vaccine reactions or not - with 9,995 vaccine adverse event mentions and 10,082 non-vaccine adverse event mentions. A larger but imbalanced dataset of 83,891 records is also available, it contains the same records as the balanced dataset plus all remaining non-vaccine adverse event mentions which were excluded when creating the balanced dataset.

Based on Twitter agreements the IDs of the datasets can be published together with the labels, so with the agreement of Twitter and provided a researcher is able to use the IDs to download the tweets and intends to use them for non-commercial research purposes, this dataset can be published for any other researcher to use in this area or any similar domain.

8.1.4 Research Question 3

What taxonomy of vaccine-related social media posts can be defined to enable classification of vaccine-related Twitter posts?

A taxonomy was derived after evaluating topic models and by examining the data — see Section 5.4 for a complete description. The taxonomy, presented in Table 19, proved useful

when analysing how the two stages of topic modelling focussed data coming from a VAEM topic and a group of topics that had similar concerns and language. This is discussed in Section 7.2.1 of the Evaluation chapter — which shows the changing distribution of the taxonomy topics in samples of the tweets over the topic modelling process, culminating in a subset where VAEM was proportionally the largest topic, and the spread of topics was reduced to mostly VAEM, discussions, and some non-adverse reaction personal health mentions.

The taxonomy is generally useful to anyone wanting to understand the topics of vaccine-related tweets. The greatest number of posts can be characterised as general discussions about vaccines. Interestingly, when it came to emphatically anti-vaccination vs. pro-vaccination discussions, the pro-vaccination group was significantly larger. News and research articles were also a substantial proportion of the texts. However, the taxonomy is likely to change over time — it was created before COVID-19 and *coronavirus* was not even a detectable subject; the research is concluding just as the first vaccines for the SARS-COV-2 virus are being tested.

8.2 The research contribution

This section discusses the contribution of the research in relation to issues that were examined in the literature review. These are that:

- Vaccines differ from drugs, and adverse events following immunization (AEFI) differ from adverse drug reactions (ADR)
- Few studies exist about social media monitoring of AEFI
- The studies that do exist focus on finding severe known reactions, and treat AEFI detection similarly to ADR detection
- Deep learning (DL) is increasingly used in social media classification tasks
- DL requires large quantities of labelled data
- There is little or no published AEFI-dedicated social media data

The literature review and the domain expert have both established that the AEFI detection in social media should not be treated like ADR, in that vaccines share many common components and reactions to them are mostly expected to be commonly experienced effects, and that they apply to a great number of mostly healthy individuals. This contrasts with the relatively few who are taking medication for illnesses, each of which have distinctive potential adverse drug reactions. Therefore, rather than trying to locate specific mentions of vaccines and known and severe adverse reactions, the research aims to find most of these conversations, no matter what their significance might be. The research introduced the term “Vaccine Adverse Event

Mentions" (VAEM) to describe these posts – which have the characteristic of being *any* adverse event mention in relation to a vaccine, rather than a known or particularly severe one. To the human observer, the most frequent feature of these posts is a declarative description of health effects in relation to an external event - similarities are observed in *how* the stories are being told as well as what they are saying. Therefore, the author took the approach of evaluating NLP tools for their ability to detect the commonalities in these conversations.

Topic modelling was first used because it can be unsupervised and not require labels, but the author used a very small set of labelled posts to score the models and identify those that most clearly discovered the similarities in the labelled VAEM posts. These models identified VAEM and other similar conversations as a distinct topic, contrasted to other topics such as pro- and anti-vaccination sentiments or news reports. One of the biggest problems with obtaining data for classification is labelling of large enough datasets, and this task can be eased if the subject of interest is proportionally well represented in the data at hand. The posts identified by topic models as most likely to contain VAEM provided this quality and quantity of data and simplified the labelling task, and so topic modelling became the first component of a pipeline to identify VAEM.

The use of a small cohort of labels to score the topic models' ability to identify texts of interest was particularly effective and simplified the evaluation of topic models – this has been adopted into the author's NLP practice and is recommended to other practitioners.

Additionally, the resulting relatively homogeneous labelled data allowed language model–based deep learning classifiers to be more effective, as they could be fine-tuned to learn the nuances that distinguish VAEM — compared with needing to understand many features that separate multiple quite different classes, which would have been the situation had the text not been first filtered by topic modelling.

This approach has enabled the gathering and subsequent identification of most of the VAEM that existed in the Twitter data, as confirmed in Section 7.4. This capacity is an important contribution of the research, for instance compared to the paper presented by J. Wang et al. (2019), which is discussed in Section **Error! Reference source not found.**. Their paper emphasized the cost of labelling and class imbalance in obtaining suitable data for detection of adverse events following flu vaccination. The authors solved this problem by first isolating users that were known to have had a vaccination, then looking only at their tweets for known reaction mentions. By contrast, this research seeks to capture all reaction mentions regardless of their provenance, then to successively filter them to reliably obtain a very high proportion of VAEM. The resulting data is very diverse and captures a lot of potential vaccine safety

signal, useful for noting trends in vaccine reactions, and avoids the potential bias of filtering to only known vaccinated users.

The study has been able to verify that the labelled VAEM data is very suited to training deep learning models, this has been reflected in the scores that were obtained by classifiers over two phases of training, first with a smaller amount of data, then with five times more data. The data is available as a balanced labelled dataset of 20 thousand records and a larger imbalanced labelled dataset of 84 thousand records – thus beginning to fill the gap of little published data.

8.3 Limitations

There are unavoidable issues and potential biases that result from using any social media data. For instance, Twitter limits the length of the posts, and the posts themselves are subject to community guidelines and mores. The Twitter free streaming API also limits access to the data. Differences in social media platforms translate to differences in users and messaging. Additionally, if data is collected over a short period then it may not represent general trends, or it may fail to capture emerging changes.

This section explores these and other limitations of the research, firstly in terms of the data that was used, and secondly in terms of the research approach and evaluation. The section also describes what has been done to mitigate some of the potential biases that were encountered in the research.

Data source: There is a limitation in the use of only Twitter as a data source for the study. While this research examined Reddit as a possible data source, it was not found to be useful for the volume and kind of posts that were needed, and Facebook data was considered but dismissed because of access problems. Twitter was found to contain the type of posts this study set out to detect, but after settling on Twitter, no further research was conducted using other platforms.

Data collection period: The data collection spanned roughly a year, and so included some potential trend patterns during the influenza seasons. However, a longer-term data collection would be better for this kind of analysis. Regarding the quantity of data, although the initial data collection over 6 months was adequate for topic modelling and an initial round of classification, it also proved to be insufficient for getting the best results from classifiers. The total data collected over a year was required to properly train and evaluate the classifiers.

Data querying: This research took the approach of using a few broad search terms, mainly around the word “vaccine” but also included “flu shot” related terms – the reasoning for using these keywords is elaborated in Section 4.1.1. These search terms constrained the data, but they

were not highly specific, and by not including a range of vaccine-specific terms it is likely that the search missed some posts. Despite these potential issues, the resulting data *did* contain sufficient VAEM examples for the study. Furthermore, those examples were a small proportion of all the data and isolating them provided the challenge that motivated this study — so the data has been suitable for the study's purpose.

As the search terms that are used determine the data that gets retrieved, and appear in the collected data, there are inevitable biases in the data towards those terms. However, the author found that as the search terms were quite general, they had no particular significance in the downstream classification of the text, and instead words that were not specified emerged as distinctive. This can be seen in the comparison of word significance in Figure 40, where although the term "flu shot" emerges as having the most significance, and indeed reflects the importance of that term to obtain VAEM-like texts, the figure shows that this term is considered significant by both the positive and negative classes. Therefore, it cannot be used for deciding on the class. Words that were *not* included in the Twitter search are much more significant for identifying VAEM – for instance "sore" and "arm" appear to be significant just for the positive class.

Data cleaning: To ensure that repeated messages did not predominate, duplicate tweets were removed, based both on tweet ID and on the text of the tweets, and retweets were also ignored. It is possible that doing so reduced the ability to measure the Twitter community's perception of the importance of a tweet, but such analysis was not a part of this research. Additionally, very short tweets were removed, and those that had significant repetition – this was done to try to decrease meaningless texts — the author believes that the benefit of doing so outweighed any potential negative cost.

Data aggregation and reduction: The research approach consisted of using topic modelling to significantly reduce the amount of data that needed to be handled by downstream classifiers, by extracting just the data from the best VAEM topic, which contained virtually all the VAEM plus similar personal health mentions and discussions. This could be characterised as creating a biased dataset, but it was intentional. The author judged that classifiers would do well at a binary classification task over a simplified dataset. An alternative approach of training multi-class classifiers to handle *all* the data was considered as unlikely to perform as well and would have also made labelling extremely costly. Reducing the dataset via topic modelling considerably eased the task of labelling — over 85% of the unwanted non-VAEM data was discarded by this filtering process. The evaluation of the results in Chapter 7 indicate that the data filtering strategy was effective.

Datasets: The labelled datasets that were produced for the classification, and which will be made available, were those data that were passed through topic modelling. Around 81% of the data was obtained from the first stage of topic modelling, the remaining 19% came from the top 3 topics of the second stage of topic modelling – see Chapter 4. Therefore, this data represents what the DMM topic models identified as VAEM and similar posts and does not represent the entire range of original data obtained by the Twitter API search. Furthermore, classifiers were trained with subsequently balanced datasets, so balanced datasets do not represent the extracted topic model data. What the datasets *do* represent is a very large range of VAEM and similar posts, which is like any other dataset that is focussed on discerning a signal within mostly homogenous data.

Metrics and Metric Selections: In this study, metrics were used when assessing the topic models and the classifiers during their training processes, then various sampling and counting were used to evaluate the effectiveness of these (and the entire model) for isolating VAEM.

The metric used for assessing topic models and classifiers were the standard measures of precision, recall, and F1-score, which count the proportions of labelled VAEM against non-VAEM. However, the topic model version of F1-scoring was not based on fully labelled data, but used a small set of 1,400 labelled records, and these records were in a dataset of several hundred thousand. Also, F1-scoring was only done for the topic that best concentrated the VAEM label. Recall was in terms of labelled VAEM in that topic against all labelled VAEM, and precision was calculated based on VAEM in that topic against all other labels in the topic – so these were proxies of an F1-Scoring system and served only to help identify the best topic models. Consequently, the scoring cannot be used as an accurate measure of the topic model or a comparative measure against any other topic models. However, when the topic models were later assessed for their effectiveness, then the subsequent labelling for classification was useful, as were samples of the excluded and unlabelled data. For instance, to verify that the best topic correctly identified VAEM, 15,000 tweets were sampled to see if the discarded tweets contained any VAEM, and it was found that only 0.04% of VAEM was likely lost – see Section 7.2.1.

The classifiers were trained on balanced datasets derived the data from the best topic model and used standard F1-Scoring, achieving F1-Scores of over 0.9. The combined effect of topic modelling and classification is estimated, by using samples and counting, to have obtained at least 90% of the VAEM – see Section 7.4.

Assessment and Interpretation: The primary goal of proving that social media can contain a significant amount of VAEM posts has been met through this approach. Extensive assessment

using counts, as discussed above, have proven this. However, this must be stated with the caveat that the assessment applies to the Twitter data gathered at the time of the study, and since the evaluation was restricted to just the approach used and that other techniques, such as multi-class classification, have not been attempted — there is no ability to compare the approach's effectiveness with others.

A potential issue with the classifiers is that they were trained on balanced data and were only tested on two reasonably small test datasets, one imbalanced dataset of 614 records, and another balanced of 828 records. They worked well in this context, but additional testing on subsequent and larger test datasets would have helped to reinforce the evaluation. However, this does not really detract from the observation that the research approach has been very successful for identifying VAEM. Verifying classifier effectiveness is only important for establishing that the data that was derived from the topic modelling process is amenable to accurate classification, the degree of accuracy is not highly important for the study.

While the evaluation has verified that the research established that there are highly useful techniques for extracting VAEM from the Twitter posts that were collected throughout 2018 and 2019, the research approach was specifically engineered to solutions that exploited distinctive features of the dataset. This is mainly the kind of language found in the generally short VAEM tweets, which describe various physical discomforts and potential reactions in relation to recent vaccinations. The orientation of the research was a practical exploration of applicable techniques for isolating VAEM, but the topic models and classifiers that were developed were specifically tuned to getting the best result from the Twitter data and were evaluated in this context. Therefore, while the approach — which has been formalised in Section 3.5 as the VAEM-Mine method — is transferrable to other similar problems, the models themselves are not designed to be generalizable enough to be applied elsewhere. Also, the VAEM-Mine method works very well for the assigned task, but since the author has not yet applied it to any other similar task, there is no evaluation of its transferability.

A small comparison study was conducted to see if the VAEM trends in the data aligned with the known flu season reporting in Victoria, Australia – see Appendix H. Although this did show alignment, there were very few Victorian-specific tweets obtainable, which highlights the problem with the Twitter data, that it is more likely to be useful in the context of the major users of the platform, which are those in the United States. Twitter may only be useful in the Australian context if the full data stream, including location specific data, is accessed.

The purpose of the study was to establish that social media could provide useful data for the task of signal detection, by providing a means of identifying the posts that contain the vaccine

adverse event mentions (which potentially contain signals). The domain expert has verified that the data does indeed provide the information required for signal detection, which would be performed by epidemiologists examining the extracted texts.

8.4 Future research

The VAEM-Mine method the research developed for detecting vaccine adverse event mentions has helped to establish that social media is indeed a prospective additional resource for vaccine safety monitoring. However, the approach needs to be implemented into a working application to realise its potential, and it also has scope for other applications, and for improvement. The following discussion explains these points.

Application of the research for vaccine safety monitoring

Vaccine safety monitoring authorities including the Therapeutic Goods Administration (TGA) and Adverse Events Following Immunisation – Clinical Assessment Network (AEFI-CAN) can benefit from the results of this research. This research potentially adds an additional active surveillance modality to complement existing passive (spontaneous) reporting systems and active surveillance systems. Currently, active systems are mostly general practice-based surveillance which utilise automated SMS to solicit reporting of AEFI experienced by recently vaccinated patients; social media-based self-reporting is an additional unsolicited surveillance source. The research findings need to be incorporated into working processes or applications that monitor social media streams for vaccine adverse event mentions. Their data would need to be assessed continuously to detect changes in trends, or specific adverse events of special interest (Law & Sturkenboom, 2020), that could indicate an emerging vaccine safety signal.

Adapting the approach to similar problems

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 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 VAEM in relation to vaccines and treatments as they are generally released.

Verifying the effectiveness of topic modelling in the NLP pipeline

The VAEM-Mine method includes a component of filtering through topic modelling, which means that data preparation for topic modelling and the subsequent topic inference needs to be built into the NLP pipeline. A future investigation needs to evaluate the trained classifier against unfiltered data, to see how well it performs in relegating previously unseen non-VAEM text patterns to the non-VAEM class. Additionally, we should compare our classifier trained on filtered data with a classifier that is trained to handle all the incoming data by separating texts into multiple classes.

Improving the topic modelling approach

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. Much more can be done in the topic modelling area, and it would be beneficial to further investigate recent work by other researchers. For instance: guided LDA (Jagarlamudi et al., 2012) might be used to teach a topic model to converge to a topic containing the most likely VAEM posts, whereas the approach used in this research was to use F1-scoring to identify when a topic model naturally placed texts of interest into a topic. The combination of the distinctiveness and similarity of VAEM facilitated this research approach but incorporating additional customized steps into the pipeline might be required to improve the approach — for instance using embeddings similarity scores to focus the initial identification of texts of interest.

In conclusion, this research has been instructive regarding the capabilities of NLP technologies, it has answered the research questions it was tasked with, and it has opened future directions, both for the research and the author, and it is hoped that much more will come as a result.

References

Abbott, T., & Morrison, S. (2015). No Jab - No Pay for Child Care. *Parliament of Australia*, 58–60. http://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id%3A%22media%2Fpressrel%2F3770236%22

After vaccination | The Australian Immunisation Handbook. (2021). https://immunisationhandbook.health.gov.au/vaccination-procedures/after-vaccination

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Almenoff, J., Tonning, J. M., Gould, A. L., Szarfman, A., Hauben, M., Ouellet-Hellstrom, R., Ball, R., Hornbuckle, K., Walsh, L., Yee, C., Sacks, S. T., Yuen, N., Patadia, V., Blum, M., Johnston, M., Gerrits, C., Seifert, H., & LaCroix, K. (2005). Perspectives on the use of data mining in pharmacovigilance. *Drug Safety*, *28*(11), 981–1007. https://doi.org/10.2165/00002018-200528110-00002

Aphinyanaphongs, Y., Brown, D. P., Langone, N. Y. U., & Krebs, P. (2016). *Text classification for Automatic Detection of E-Cigarette Use and Use for Smoking Cessation From Twitter : a Feasibility Pilot*. 480–491.

Armstrong, P. K., Dowse, G. K., Effler, P. V., Carcione, D., Blyth, C. C., Richmond, P. C., Geelhoed, G. C., Mascaro, F., Scully, M., & Weeramanthri, T. S. (2011). Epidemiological study of severe febrile reactions in young children in Western Australia caused by a 2010 trivalent inactivated influenza vaccine. *BMJ Open*, *1*(1), e000016–e000016. https://doi.org/10.1136/bmjopen-2010-000016

Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*, 507–512.

Baker, M. A., Nguyen, M., Cole, D. V., Lee, G. M., & Lieu, T. A. (2013). Post-licensure rapid immunization safety monitoring program (PRISM) data characterization. *Vaccine*, *31*(S10), K98–K112. https://doi.org/10.1016/j.vaccine.2013.04.088

Baldwin, T., Marneffe, M. C. De, Han, B., Kim, Y., & Ritter, A. (2015). *Shared Tasks of the 2015 Workshop on Noisy User-generated Text : Twitter Lexical Normalization and Named Entity Recognition*. 126–135.

Barde, B. V., & Bainwad, A. M. (2017). *An Overview of Topic Modeling Methods and Tools*. 745–750.

Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, *66*, 125–

136. https://doi.org/10.1016/j.future.2016.06.032

Benton, A., Paul, M. J., Hancock, B., & Dredze, M. (2016). Collective supervision of topic models for predicting surveys with social media. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, 2892–2898.

Bicalho, P., Pita, M., Pedrosa, G., Lacerda, A., & Pappa, G. L. (2017). A general framework to expand short text for topic modeling. *Information Sciences*, *393*, 66–81. https://doi.org/10.1016/j.ins.2017.02.007

Blei, D. M., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). *Fast unfolding of communities in large networks*. https://doi.org/10.1088/1742-5468/2008/10/P10008

Budhiraja, S., & Akinapelli, R. (2010). Pharmacovigilance in vaccines. *Indian Journal of Pharmacology*, *42*(2), 117.

Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency Situation Awareness from Twitter for crisis management. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion*, 695–698. https://doi.org/10.1145/2187980.2188183

Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, 288--296. https://doi.org/10.1.1.100.1089

Chang, T.-Y., & Chen, Y.-N. (2019). What does this word mean? Explaining contextualized embeddings with natural language definition. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6066–6072.

Chen, R. T. (1999). Vaccine risks: Real, perceived and unknown. *Vaccine*, *17*(SUPPL. 3), 41–46. https://doi.org/10.1016/S0264-410X(99)00292-3

Chen, R. T., Rastogi, S. C., Mullen, J. R., Hayes, S. W., Cochi, S. L., Donlon, J. A., & Wassilak, S. G. (1994). The vaccine adverse event reporting system (VAERS). *Vaccine*, *12*(6), 542–550. https://doi.org/10.1016/0264-410X(94)90315-8

Chen, R. T., Shimabukuro, T. T., Martin, D. B., Zuber, P. L. F., Weibel, D. M., & Sturkenboom, M. (2015). Enhancing vaccine safety capacity globally: A lifecycle perspective. *Vaccine*, *33*, D46–D54. https://doi.org/10.1016/j.vaccine.2015.06.073

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). *BTM : Topic Modeling over Short Texts*. *26*(12), 2928–2941.

Choi, B. C. K. (2012). *The Past , Present , and Future of Public Health Surveillance*. *2012*(Table 1).

Choi, D., Matni, Z., & Shah, C. (2016). What social media data should i use in my research?: A comparative analysis of twitter, youtube, reddit, and the new york times comments. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–6. https://doi.org/10.1002/pra2.2016.14505301151

Chowdhury, S., Zhang, C., & Yu, P. S. (2018). *Multi-Task Pharmacovigilance Mining from Social Media Posts*. *1*, 1–10. https://doi.org/10.1145/3178876.3186053

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *American Journal of Tropical Medicine and Hygiene*, *86*(1), 39–45. https://doi.org/10.4269/ajtmh.2012.11-0597

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014. *ArXiv Preprint ArXiv:1412.3555*.

Clothier, H. J., Crawford, N. W., Kempe, A., & Buttery, J. P. (2011). Surveillance of adverse events following immunisation: the model of SAEFVIC, Victoria. *Communicable Diseases Intelligence*, *35*(4), 294–298.

Clothier, H. J., Lawrie, J., Russell, M. A., Kelly, H., & Buttery, J. P. (2019). Early signal detection of adverse events following influenza vaccination using proportional reporting ratio, Victoria, Australia. *PLoS ONE*, *14*(11), 1–17. https://doi.org/10.1371/journal.pone.0224702

Clothier, H. J., Selvaraj, G., Easton, M. L., Lewis, G., Crawford, N. W., & Buttery, J. P. (2014). Consumer reporting of adverse events following immunization. *Human Vaccines and Immunotherapeutics*, *10*(12), 3726–3730. https://doi.org/10.4161/hv.34369

Cocos, A., Fiks, A. G., & Masino, A. J. (2017). Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association*, *24*(4), 813–821. https://doi.org/10.1093/jamia/ocw180

Conway, M., Hu, M., & Chapman, W. W. (2019). Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. *Yearbook of Medical Informatics*, *28*(1), 208.

Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder

in twitter. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 579–582.

Council for International Organizations of Medical Sciences & WHO. (2012). Definition and application of terms for vaccine pharmacovigilance: report of CIOMS/WHO Working Group on Vaccine Pharmacovigilance. *Geneva*, 39–40. http://www.who.int/vaccine_safety/initiative/tools/CIOMS_report_WG_vaccine.pdf

Crawford, N. W., Clothier, H., Hodgson, K., Selvaraj, G., Easton, M. L., & Buttery, J. P. (2014). Active surveillance for adverse events following immunization. *Expert Review of Vaccines*, *13*(2), 265–276. https://doi.org/10.1586/14760584.2014.866895

Cutrone, R., Lednicky, J., Dunn, G., Rizzo, P., Bocchetta, M., Chumakov, K., Minor, P., & Carbone, M. (2005). Some oral poliovirus vaccines were contaminated with infectious SV40 after 1961. *Cancer Research*, *65*(22), 10273–10279. https://doi.org/10.1158/0008-5472.CAN-05-2028

Dai Nguyen, T. (2019). *Rich and Scalable Models for Text*. University of Maryland, College Park.

Daume III, H., & Marcu, D. (2006). Domain Adaptation for Statistical Classiers. *Journal of Artificial Intelligence Research*, *26*, 101–126. https://doi.org/10.1613/jair.1872

Deiner, M. S., Fathy, C., Kim, J., Niemeyer, K., Ramirez, D., Ackley, S. F., Liu, F., Lietman, T. M., & Porco, T. C. (2017). Facebook and Twitter vaccine sentiment in response to measles outbreaks. *Health Informatics Journal*, 146045821774072. https://doi.org/10.1177/1460458217740723

Delir Haghighi, P., Kang, Y.-B., Buchbinder, R., Burstein, F., & Whittle, S. (2017). Investigating Subjective Experience and the Influence of Weather Among Individuals With Fibromyalgia: A Content Analysis of Twitter. *JMIR Public Health and Surveillance*, *3*(1), e4. https://doi.org/10.2196/publichealth.6344

Deng, L. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Dol, J., Tutelman, P. R., Chambers, C. T., Barwick, M., Drake, E. K., Parker, J. A., Parker, R., Benchimol, E. I., George, R. B., & Witteman, H. O. (2019). Health researchers' use of social media: Scoping review. *Journal of Medical Internet Research*, *21*(11), 1–12. https://doi.org/10.2196/13687

Du, J., Xu, J., Song, H. Y., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making*, *17*(Suppl 2). https://doi.org/10.1186/s12911-017-0469-6

Edo-Osagie, O., De La Iglesia, B., Lake, I., & Edeghere, O. (2020). A scoping review of the use of Twitter for public health research. *Computers in Biology and Medicine*, 103770.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., & Seligman, M. E. P. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*, *26*(2), 159–169. https://doi.org/10.1177/0956797614557867

Evans, H. P., Cooper, A., Williams, H., & Carson-Stevens, A. (2016). Improving the safety of vaccine delivery. *Human Vaccines & Immunotherapeutics*, *12*(5), 1280–1281. https://doi.org/10.1080/21645515.2015.1137404

Farzindar, A., & Inkpen, D. (2015). Natural Language Processing for Social Media. *Synthesis Lectures on Human Language Technologies*, *8*(2), 1–166. https://doi.org/10.1007/s13398-014-0173-7.2

Freed, G. L., Clark, S. J., Butchart, A. T., Singer, D. C., & Davis, M. M. (2010). Parental Vaccine Safety Concerns in 2009. *Pediatrics*, *125*(4), 654–659. https://doi.org/10.1542/peds.2009-1962

Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., & Dasgupta, N. (2014). Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter. *Drug Safety*, *37*(5), 343–350. https://doi.org/10.1007/s40264-014-0155-x

Full Reddit Submission Corpus. (2015). https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/

Gephi-The Open Graph Viz Platform. (n.d.). https://gephi.org/

Ghosh, D., & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science*, *40*(2), 90–102. https://doi.org/10.1080/15230406.2013.776210

Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of Medical Informatics*, *26*(01), 214–227.

https://doi.org/10.15265/IY-2017-029

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2008). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868.

Griffin, M. R., Braun, M. M., & Bart, K. J. (2009). What should an ideal vaccine postlicensure safety system be? *American Journal of Public Health*, *99*(SUPPL. 2), 345–350. https://doi.org/10.2105/AJPH.2008.143081

Gu, Y., Qian, Z., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*, 321–342. https://doi.org/10.1016/j.trc.2016.02.011

Gualtieri, L. N. (2009). The doctor as the second opinion and the internet as the first. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 2489–2498).

Gupta, A., & Katarya, R. (2020). Social media based surveillance systems for healthcare using machine learning: A systematic review. *Journal of Biomedical Informatics*, *108*(April 2019), 103500. https://doi.org/10.1016/j.jbi.2020.103500

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hardt, K., Schmidt-Ott, R., Glismann, S., Adegbola, R., & Meurice, F. (2013). Sustaining Vaccine Confidence in the 21st Century. *Vaccines*, *1*(3), 204–224. https://doi.org/10.3390/vaccines1030204

Härmark, L., & Van Grootheest, A. C. (2008). Pharmacovigilance: Methods, recent developments and future perspectives. *European Journal of Clinical Pharmacology*, *64*(8), 743–752. https://doi.org/10.1007/s00228-008-0475-9

Harpaz, R., DuMochel, W., & Shah, N. H. (2016). Big data and adverse drug reaction detection. *Clinical Pharmacology and Therapeutics*, *99*(3), 268–270. https://doi.org/10.1002/cpt.302

Heaton, J. (2016). *An Empirical Analysis of Feature Engineering for Predictive Modeling*. 0–5. https://doi.org/10.1109/SECON.2016.7506650

Hinrichsen, V. L., Kruskal, B., O'Brien, M. A., Lieu, T. A., Platt, R., & For, M. S. C. (2007). Using Electronic Medical Records to Enhance Detection and Reporting of Vaccine Adverse Events. *Journal of the American Medical Informations Association*, *14*(6), 731–735. https://doi.org/10.1197/jamia.M2232.Introduction

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

Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online Learning for latent Dirichlet allocation

(Supplementary Material). *Nature*, 1–9. https://doi.org/10.1.1.187.1883

Honnibal, M. (2017). *Spacy*.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, *1*, 328–339.

Huang, X., Smith, M. C., Jamison, A. M., Broniatowski, D. A., Dredze, M., Quinn, S. C., Cai, J., & Paul, M. J. (2018). *Can online self-reports assist in real- time identification of influenza vaccination uptake ? A cross-sectional study of influenza vaccine-related tweets in the USA , 2013 – 2017*. 1–7. https://doi.org/10.1136/bmjopen-2018-024018

Huang, X., Smith, M. C., Paul, M. J., Ryzhkov, D., Quinn, S. C., Broniatowski, D. A., & Dredze, M. (2017). Examining Patterns of Influenza Vaccination in Social Media. *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 1–5. http://www.cs.jhu.edu/~mdredze/publications/2017_w3phi_vaccines.pdf

Huynh, T., He, Y., Willis, A., & Rüger, S. (2016). Adverse Drug Reaction Classification With Deep Neural Networks. *Proceedings of COLING 2016: Technical Papers, COLING*, 877–887.

Isaacs, D., Lawrence, G., Boyd, I., Ronaldson, K., & Mcewen, J. (2005). *Reporting of adverse events following immunization in Australia*. *December 2004*, 163–166.

Iyer, A., Joshi, A., Karimi, S., Sparks, R., & Paris, C. (2019). *Figurative Usage Detection of Symptom Words to Improve Personal Health Mention Detection*. *November 2018*, 1142–1147. https://doi.org/10.18653/v1/p19-1108

Ja, B., Sa, H., Vaudry W, Bj, L., & Dw, S. (2014). The Canadian Immunization Monitoring Program, ACTive (IMPACT): Active surveillance for vaccine adverse events and vaccine-preventable diseases. *Public Health Agency of Canada*, *40*(4), 41–44. http://www.phacaspc.gc.ca/publicat/ccdr-rmtc/14vol40/dr-rm40s-3/comment-d-eng.php

Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 204–213.

Jiang, K., Feng, S., Calix, R. A., & Bernard, G. R. (2019). Assessment of Word Embedding Techniques for Identification of Personal Experience Tweets Pertaining. *Precision Health and Medicine: A Digital Revolution in Healthcare*, *843*, 45.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*.

Jonnagaddala, J., Jue, T. R., & Dai, H.-J. (2016). *Binary classification of Twitter posts for*

adverse drug reactions.

Joshi, A., Dai, X., Karimi, S., & Macintyre, C. R. (2018). *Shot Or Not : Comparison of NLP Approaches for Vaccination Behaviour Shot Or Not : Comparison of NLP Approaches for Vaccination Behaviour Detection*. *October*.

Joshi, A., Karimi, S., Sparks, R., Paris, C., & Macintyre, C. R. (2019). Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective. *ACM Computing Surveys (CSUR)*, *52*(6), 1–19.

Joshi, A., Sparks, R., McHugh, J., Karimi, S., Paris, C., & MacIntyre, C. R. (2020). Harnessing Tweets for Early Detection of an Acute Disease Event. *Epidemiology (Cambridge, Mass.)*, *31*(1), 90–97. https://doi.org/10.1097/EDE.0000000000001133

Jurafsky, D., & Martin, J. H. (2007). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Computational Linguistics*, *26*(4), 638–641. https://doi.org/10.1162/089120100750105975

Karapetiantz, P., Bellet, F., Audeh, B., Lardon, J., & Meyer, J. C. (2018). *Descriptions of Adverse Drug Reactions Are Less Informative in Forums Than in the French Pharmacovigilance Database but Provide More Unexpected Reactions*. *9*(May), 1–11. https://doi.org/10.3389/fphar.2018.00439

Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., & Paris, C. (2015). Text and Data Mining Techniques in Adverse Drug Reaction Detection. *ACM Computing Surveys*, *47*(4), 1–39. https://doi.org/10.1145/2719920

Karisani, P., & Agichtein, E. (2018). *Did You Really Just Have a Heart Attack? Towards Robust Detection of Personal Health Mentions in Social Media*. 137–146. http://arxiv.org/abs/1802.09130

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, *52*(4). https://doi.org/10.1145/3343440

Khademi, S., & Haghighi, P. D. (2019). *Topic Modelling for Vaccine Safety Signal Detection Vaccines Importance*.

Khademi, S., Haghighi, P. D., Lewis, P., Burstein, F., & Palmer, C. (2015). *Intelligent audit code generation from free text in the context of neurosurgery*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1746–1751. https://doi.org/10.3115/v1/d14-1181

Klein, A. Z., Alimova, I., Flores, I., Magge, A., Miftahutdinov, Z., Minard, A.-L., O'Connor, K., Sarker, A., Tutubalina, E., Weissenbacher, D., & Gonzalez-Hernandez, G. (2020). Overview of the Fifth Social Media Mining for Health (SMM4H) Shared Tasks at COLING 2020. *Proceedings of the Fifth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. https://doi.org/10.18653/v1/w19-3203

Kohl, K. S., Gidudu, J., Bonhoeffer, J., Braun, M. M., Buettcher, M., Chen, R. T., Drammeh, B., Duclos, P., Heijbel, H., Heininger, U., Hummelman, E., Jefferson, T., Keller-Stanislawski, B., Loupi, E., & Marcy, S. M. (2007). The development of standardized case definitions and guidelines for adverse events following immunization. *Vaccine*, *25*(31), 5671–5674. https://doi.org/10.1016/j.vaccine.2007.02.063

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, *10*(4), 1–68. https://doi.org/10.3390/info10040150

Krska, J., Anderson, C., Murphy, E., & Avery, A. J. (2011). How patient reporters identify adverse drug reactions: A qualitative study of reporting via the UK yellow card scheme. In *Drug Safety* (Vol. 34, Issue 5, pp. 429–436). https://doi.org/10.2165/11589320-000000000-00000

Lama, Y., Hu, D., Jamison, A., Quinn, S. C., & Broniatowski, D. A. (2019). Characterizing Trends in Human Papillomavirus Vaccine Discourse on Reddit (2007-2015): An Observational Study. *JMIR Public Health and Surveillance*, *5*(1), e12480. https://doi.org/10.2196/12480

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on Twitter. *Proceedings of NAACL-HLT 2013*, *June*, 789–795.

Lample, G., & Conneau, A. (2019). *Cross-lingual Language Model Pretraining*. http://arxiv.org/abs/1901.07291

Lampos, V., Zou, B., & Cox, I. J. (2017). Enhancing feature selection using word embeddings: The case of flu surveillance. *Www '17*, *Ili*, 695–704. https://doi.org/10.1145/3038912.3052622

Lantos, J. D., Jackson, M. A., Opel, D. J., Marcuse, E. K., Myers, A. L., & Connelly, B. L. (2010). Controversies in Vaccine Mandates. *Current Problems in Pediatric and Adolescent Health Care*, *40*(3), 38–58. https://doi.org/10.1016/j.cppeds.2010.01.003

Lardon, J., Abdellaoui, R., Bellet, F., Asfari, H., Souvignet, J., Texier, N., Jaulent, M. C., Beyens, M. N., Burgun, A., & Bousquet, C. (2015). Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*,

17(7), 1–16. https://doi.org/10.2196/jmir.4304

Lardon, J., Bellet, F., Aboukhamis, R., Asfari, H., Souvignet, J., Jaulent, M.-C., Beyens, M.-N., Lillo-LeLouët, A., & Bousquet, C. (2018). Evaluating Twitter as a complementary data source for pharmacovigilance. *Expert Opinion on Drug Safety*, *17*(8), 763–774.

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., & Ratzan, S. (2011). Addressing the vaccine confidence gap. *The Lancet*, *378*(9790), 526–535. https://doi.org/10.1016/S0140-6736(11)60678-8

Larson, H. J., Smith, D. M. D., Paterson, P., Cumming, M., Eckersberger, E., Freifeld, C. C., Ghinai, I., Jarrett, C., Paushter, L., Brownstein, J. S., & Madoff, L. C. (2013). Measuring vaccine confidence: Analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines. *The Lancet Infectious Diseases*, *13*(7), 606–613. https://doi.org/10.1016/S1473-3099(13)70108-7

Latent Dirichlet Allocation via Mallet. (2020). https://radimrehurek.com/gensim/models/wrappers/ldamallet.html

Law, B., & Sturkenboom, M. (2020). D2. 3 priority list of adverse events of special interest: COVID-19. *Oslo, NO: CEPI*.

Lazarus, R., Klompas, M., Campion, F. X., McNabb, S. J. N., Hou, X., Daniel, J., Haney, G., DeMaria, A., Lenert, L., & Platt, R. (2009). Electronic Support for Public Health: Validated Case Finding and Reporting for Notifiable Diseases Using Electronic Medical Data. *Journal of the American Medical Informatics Association*, *16*(1), 18–24. https://doi.org/10.1197/jamia.M2848

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

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

Lee, K., Qadir, A., Hasan, S. A., Datla, V., Prakash, A., Liu, J., & Farri, O. (2017). Adverse Drug Event Detection in Tweets with Semi-Supervised Convolutional Neural Networks. *Proceedings of the 26th International Conference on World Wide Web - WWW '17*, 705–714. https://doi.org/10.1145/3038912.3052671

Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2017). Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems (TOIS)*, *36*(2), 1–30.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016). Topic modeling for short texts with

auxiliary word embeddings. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165–174.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. *1*. http://arxiv.org/abs/1907.11692

Lopalco, P. L., Johansen, K., Ciancio, B., De Carvalho Gomes, H., Kramarz, P., & Giesecke, J. (2010). Monitoring and assessing vaccine safety: a European perspective. *Expert Review of Vaccines*, *9*(4), 371–380. https://doi.org/10.1586/erv.10.20

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., & Zhang, G. (2015). Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, *80*, 14–23. https://doi.org/10.1016/j.knosys.2015.01.010

MacDonald, N. E., Harmon, S., Dube, E., Steenbeek, A., Crowcroft, N., Opel, D. J., Faour, D., Leask, J., & Butler, R. (2018). Mandatory infant & childhood immunization: Rationales, issues and knowledge gaps. *Vaccine*, *36*(39), 5811–5818.

Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Marshall, V., & Baylor, N. W. (2011). Food and Drug Administration regulation and evaluation of vaccines. *Pediatrics*, *127*(SUPPL. 1). https://doi.org/10.1542/peds.2010-1722E

Martin, F., & Johnson, M. (2015). More efficient topic modelling through a noun only approach. *Proceedings of the Australasian Language Technology Association Workshop 2015*, 111–115.

Mazarura, J., & Waal, A. De. (2016). *A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text*. 1–6. https://doi.org/10.1109/RoboMech.2016.7813155

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. *Advances in Neural Information Processing Systems*, 121–128.

Mccallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. citeulike-article-id:1062263

McGough, S. F., Brownstein, J. S., Hawkins, J. B., & Santillana, M. (2017). Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLoS Neglected Tropical Diseases*, *11*(1), 1–15. https://doi.org/10.1371/journal.pntd.0005295

McNeil, M. M., Gee, J., Weintraub, E. S., Belongia, E. A., Lee, G. M., Glanz, J. M., Nordin, J. D., Klein, N. P., Baxter, R., Naleway, A. L., Jackson, L. A., Omer, S. B., Jacobsen, S. J., & DeStefano, F. (2014). The Vaccine Safety Datalink: Successes and challenges monitoring vaccine safety. *Vaccine*, *32*(42), 5390–5398. https://doi.org/10.1016/j.vaccine.2014.07.073

Mcphillips, H., & Marcuse, E. K. (2001). *Vaccine Safety*. *April*, 95–121. https://doi.org/10.1067/mps.2001.113988

Mehta, H. (2020). *Validating medical queries with literature from pubmed using topic modelling*.

Mesfin, Y., Cheng, A. C., Enticott, J., Lawrie, J., & Buttery, J. P. (2020). Use of telephone helpline data for syndromic surveillance of adverse events following immunization in Australia: A retrospective study, 2009 to 2017. *Vaccine*, *38*(34), 5525–5531.

Mesfin, Y., Cheng, A., Lawrie, J., & Buttery, J. (2019). Use of routinely collected electronic healthcare data for postlicensure vaccine safety signal detection: a systematic review. *BMJ Global Health*, *4*(4), e001065.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*, *September*, 1045–1048.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 1–9.

Milinovich, G. J., Williams, G. M., Clements, A. C. A., & Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet Infectious Diseases*, *14*(2), 160–168. https://doi.org/10.1016/S1473-3099(13)70244-5

Milstien, J. B., Batson, A., & Wertheimer, A. I. (2015). *Vaccines and drugs: characteristics of their use to meet public health goals (English)*. *March*, 1–40. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2005/04/14/000090341_20050414151834/Rendered/PDF/320400MilstienVaccinesDrugsFinal.pdf

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mollema, L., Harmsen, I. A., Broekhuizen, E., Clijnk, R., De Melker, H., Paulussen, T., Kok,

G., Ruiter, R., & Das, E. (2015). Disease detection or public opinion reflection? content analysis of tweets, other social media, and online newspapers during the measles outbreak in the Netherlands in 2013. *Journal of Medical Internet Research*, *17*(5). https://doi.org/10.2196/jmir.3863

National centre of Immunization. (2017). *No Jab No Play, No Jab No Pay Policies*. http://www.ncirs.edu.au/consumer-resources/no-jab-no-play-no-jab-no-pay-policies/

Newman, D., Lau, J., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. ... *Language Technologies: The ...*, *June*, 100–108. https://doi.org/10.3115/1220175.1220274

Nguyen, D. Q. (2018). *jLDADMM: A Java package for the LDA and DMM topic models*. *Dmm*, 1–5. http://arxiv.org/abs/1808.03835

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*, 299–313.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359.

Nigam, K., Mccallum, A. K., Thrun, S., & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, *39*, 103–134. https://doi.org/10.1023/A:1007692713085

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, *22*(3), 671–681. https://doi.org/10.1093/jamia/ocu041

NLTK, Natural Language Toolkit. (2018). http://www.nltk.org/

Nugroho, R., Paris, C., Nepal, S., Yang, J., & Zhao, W. (2020). A survey of recent methods on deriving topics from Twitter: algorithm to evaluation. In *Knowledge and Information Systems*. Springer London. https://doi.org/10.1007/s10115-019-01429-z

O'Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International Journal of Medical Informatics*, *101*, 15–22. https://doi.org/10.1016/j.ijmedinf.2017.01.019

Ofoghi, B., Mann, M., & Verspoor, K. (2016). Towards Early Discovery of Salient Health Threats: a Social Media Emotion Classification Technique. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *21*, 504–515.

Pal, S. N., Duncombe, C., Falzon, D., & Olsson, S. (2013). WHO strategy for collecting safety data in public health programmes: Complementing spontaneous reporting systems. *Drug Safety*, *36*(2), 75–81. https://doi.org/10.1007/s40264-012-0014-6

Palogiannidi, E., Kolovou, A., Christopoulou, F., Kokkinos, F., Iosif, E., Malandrakis, N., Papageorgiou, H., Narayanan, S., & Potamianos, A. (2016). Tweester at SemEval-2016 task 4: Sentiment analysis in twitter using semantic-affective model adaptation. *SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings*, 155–163. https://doi.org/10.18653/v1/s16-1023

Paparrizos, J., White, R. W., & Horvitz, E. (2016). Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results. *Journal of Oncology Practice*, *12*(8), 737–744. https://doi.org/10.1200/JOP.2015.010504

Parrella, A., Braunack-Mayer, A., Gold, M., Marshall, H., & Baghurst, P. (2013). Healthcare providers' knowledge, experience and challenges of reporting adverse events following immunisation: a qualitative study. *BMC Health Services Research*, *13*(1), 313. https://doi.org/10.1186/1472-6963-13-313

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 8024–8035.

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 265–272. https://doi.org/10.1.1.224.9974

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0103408

Paul, M. J., & Dredze, M. (2017). Social Monitoring for Public Health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *9*(5), 1–183. https://doi.org/10.2200/S00791ED1V01Y201707ICR060

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 1756–1765. https://doi.org/10.18653/v1/P17-1161

Pimpalkhute, P., Patki, A., Nikfarjam, A., & Gonzalez, G. (2014). Phonetic spelling filter for

keyword selection in drug mention mining from social media. *AMIA Joint Summits on Translational Science Proceedings AMIA Summit on Translational Science*, 90–95. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4333687&tool=pmcentrez&rendertype=abstract

Poursabzi-Sangdeh, F., & Boyd-Graber, J. (2015). *Speeding Document Annotation with Topic Models*. 126–132. https://doi.org/10.3115/v1/n15-2017

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health-related topics on twitter an exploration of tobacco-related tweets as a test topic. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6589 LNCS*, 18–25. https://doi.org/10.1007/978-3-642-19656-0_4

Principi, N., & Esposito, S. (2016). Adverse events following immunization: real causality and myths. *Expert Opinion on Drug Safety*, *15*(6), 825–835. https://doi.org/10.1517/14740338.2016.1167869

Priya, S., Sequeira, R., Chandra, J., & Dandapat, S. K. (2019). Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media*, *9*, 17–29. https://doi.org/10.1016/j.osnem.2018.11.001

Pulendran, B., & Ahmed, R. (2011). Immunological mechanisms of vaccination. *Nature Immunology*, *12*(6), 509–517. https://doi.org/10.1038/ni.2039

pyLDAvis - Python library for interactive topic model visualization. (2018). https://github.com/bmabey/pyLDAvis

Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2019). Short text topic modeling techniques, applications, and performance: A survey. *ArXiv*, *14*(8). https://doi.org/10.1109/tkde.2020.2992485

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. *IJCAI International Joint Conference on Artificial Intelligence*, *2015-Janua*(Ijcai), 2270–2276.

Radzikowski, J., Stefanidis, A., Jacobsen, K. H., Croitoru, A., Crooks, A., & Delamater, P. L. (2016). The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health and Surveillance*, *2*(1), e1. https://doi.org/10.2196/publichealth.5059

Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21. https://doi.org/10.1109/JBHI.2016.2636665

Reddit's publicly available comment dataset. (2015).

https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

Řehůřek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora*. http://www.muni.cz/research/publications/884893

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V., & Boyd-graber, J. (2015). Beyond LDA : Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. *Proceedings of the 52nd Workshop Computational Linguistics and Clinical Psychology*, *1*(2014), 99–107. https://doi.org/10.3115/v1/W15-1212

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 399–408. https://doi.org/10.1145/2684822.2685324

Rong, J., Michalska, S., Subramani, S., Du, J., & Wang, H. (2019). Deep learning for pollen allergy surveillance from twitter in Australia. *BMC Medical Informatics and Decision Making*, *19*(1), 1–13. https://doi.org/10.1186/s12911-019-0921-x

Sadilek, A., Kautz, H., Di Prete, L., Labus, B., Portman, E., Teitel, J., & Silenzio, V. (2017). Deploying nemesis: Preventing foodborne illness by data mining social media. *AI Magazine*, *38*(1), 37–48. https://doi.org/10.1609/aimag.v38i1.2711

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, *7*(10). https://doi.org/10.1371/journal.pcbi.1002199

Salmon, D. A., & Halsey, N. A. (2016). How Vaccine Safety is Monitored. *The Vaccine Book: Second Edition*, 153–165. https://doi.org/10.1016/B978-0-12-802174-3.00008-4

Sarker, A., Chandrashekar, P., Magge, A., Cai, H., Klein, A., & Gonzalez, G. (2017). Discovering Cohorts of Pregnant Women From Social Media for Safety Surveillance and Analysis. *Journal of Medical Internet Research*, *19*(10), e361. https://doi.org/10.2196/jmir.8164

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., & Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, *54*, 202–212. https://doi.org/10.1016/j.jbi.2015.02.004

Sarker, A., & Gonzalez-Hernandez, G. (2017). Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017. *CEUR Workshop Proceedings*, *1996*, 43–48.

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug

reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, *53*, 196–207. https://doi.org/10.1016/j.jbi.2014.11.002

Sarker, A., & Gonzalez, G. (2017). A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. *Data in Brief*, *10*, 122–131. https://doi.org/10.1016/j.dib.2016.11.056

Seifert, H. A., Malik, R. E., Bhattacharya, M., Campbell, K. R., Okun, S., Pierce, C., Terkowitz, J., Rick Turner, J., Krucoff, M. W., & Powell, G. E. (2017). Enabling Social Listening for Cardiac Safety Monitoring: Proceedings from a DIA-CSRC Co-sponsored Think Tank. *American Heart Journal*, *194*, 107–115. https://doi.org/10.1016/j.ahj.2017.08.021

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1715–1725. https://doi.org/10.18653/v1/p16-1162

Settles, B. (2012). Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

Sievert, C., Shirley, K. E., & York, N. (2014). *LDAvis : A method for visualizing and interpreting topics*. 63–70.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0019467

Singh, T., Roberts, K., Cohen, T., Cobb, N., Wang, J., Fujimoto, K., & Myneni, S. (2020). Social media as a research tool (smaart) for risky behavior analytics: Methodological review. *JMIR Public Health and Surveillance*, *6*(4). https://doi.org/10.2196/21660

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). *Twitter as a Tool for Health Research : A Systematic Review*. *107*(1), 1–9. https://doi.org/10.2105/AJPH.2016.303512

Steele, R. (2011). Social media, mobile devices and sensors: Categorizing new techniques for health communication. *Proceedings of the International Conference on Sensing Technology, ICST*, 187–192. https://doi.org/10.1109/ICSensT.2011.6136960

Stokes, B. (2010). *Ministerial review into the public health response into the adverse events to the seasonal influenza vaccine. Final report to the minister for health*. *2011*(14 Jan).

Sun, X., Ren, F., & Ye, J. (2017). Trends detection of flu based on ensemble models with emotional factors from social networks. *IEEJ Transactions on Electrical and Electronic Engineering*, *12*(3), 388–396. https://doi.org/10.1002/tee.22389

Surian, D., Nguyen, D. Q., Kennedy, G., Johnson, M., Coiera, E., & Dunn, A. G. (2016).

Characterizing twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of Medical Internet Research*, *18*(8), 1–12. https://doi.org/10.2196/jmir.6045

Tan, S. S.-L., & Goonawardene, N. (2017). Internet health information seeking and the patient-physician relationship: a systematic review. *Journal of Medical Internet Research*, *19*(1), e9.

Tang, L., Bie, B., Park, S. E., & Zhi, D. (2018). Social media and outbreaks of emerging infectious diseases: A systematic review of literature. *American Journal of Infection Control*, *46*(9), 962–972. https://doi.org/10.1016/j.ajic.2018.02.010

Vaccine history timeline. (2018). https://www2.health.vic.gov.au/public-health/immunisation/immunisation-schedule-vaccine-eligibility-criteria/vaccine-history-timeline

Van Hee, C., Jacobs, G., Emmery, C., DeSmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS ONE*, *13*(10), 1–22. https://doi.org/10.1371/journal.pone.0203794

Varricchio, F., Iskander, J., Destefano, F., Ball, R., Pless, R., Braun, M. M., & Chen, R. T. (2004). Understanding vaccine safety information from the Vaccine Adverse Event Reporting System. *Pediatric Infectious Disease Journal*, *23*(4), 287–294. https://doi.org/10.1097/00006454-200404000-00002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Velardi, P., Stilo, G., Tozzi, A. E., & Gesualdo, F. (2014). Twitter mining for fine-grained syndromic surveillance. *Artificial Intelligence in Medicine*, *61*(3), 153–163. https://doi.org/10.1016/j.artmed.2014.01.002

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, *4*, 1105–1112.

Wang, C., Paisley, J., & Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. *Journal of Machine Learning Research*, *15*, 752–760.

Wang, J., & Zhao, L. (2018). *Multi-instance Domain Adaptation for Vaccine Adverse Event Detection*. 97–106. https://doi.org/10.1145/3178876.3186051

Wang, J., Zhao, L., & Ye, Y. (2019). Semi-supervised Multi-instance Interpretable Models for Flu Shot Adverse Event Detection. *Proceedings - 2018 IEEE International Conference*

on Big Data, Big Data 2018, *October*, 851–860. https://doi.org/10.1109/BigData.2018.8622434

Wang, J., Zhao, L., Ye, Y., & Zhang, Y. (2018). Adverse event detection by integrating twitter data and VAERS. *Journal of Biomedical Semantics*, *9*(1), 1–10. https://doi.org/10.1186/s13326-018-0184-y

Wang, S., & Manning, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, *July*, 90–94.

Wang, Y., Li, X., & Mo, D. Y. (2021). *Personal Health Mention Identification from Tweets Using Convolutional Neural Network*. 650–654. https://doi.org/10.1109/ieem45057.2020.9309807

Weissenbacher, D., & Gonzalez-Hernandez, G. (2019). *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. https://www.aclweb.org/anthology/W19-3200

Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M., & Gonzalez, G. (2019). Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, 21–30.

Weissenbacher, D., Sarker, A., Paul, M., & Gonzalez, G. (2018). Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, 13–16.

WHO. (2010). Guidelines for independent lot release of vaccines by regulatory authorities TRS (ECBS 2010). *Guidance*, *Ecbs*, 2–27.

WHO. (2013). *Causality assessment of an adverse event following immunization (AEFI). User manual for the revised WHO AEFI causality assessment classification*. 56. http://www.who.int/vaccine_safety/publications/gvs_aefi/en/

WHO. (2019). *Ten health issues WHO will tackle this year*. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

Win, S. S. M., & Aung, T. N. (2018). Automated text annotation for social media data during natural disasters. *Advances in Science, Technology and Engineering Systems*, *3*(2), 119–127. https://doi.org/10.25046/aj030214

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). *HuggingFace's Transformers: State-of-the-art*

Natural Language Processing. http://arxiv.org/abs/1910.03771

World Health Organization. (2002). The Importance of Pharmacovigilance - Safety Monitoring of medicinal products. *Who*, 1–52. https://doi.org/10.1002/0470853093

Yadav, P., Steinbach, M., Kumar, V., & Simon, G. (2018). Mining electronic health records (EHRs) A survey. *ACM Computing Surveys (CSUR)*, *50*(6), 1–40.

Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., Lehman, C., Buckley, J. M., Coopey, S. B., Polubriaginof, F., Garber, J. E., Smith, B. L., Gadd, M. A., Specht, M. C., Gudewicz, T. M., Guidi, A. J., Taghian, A., & Hughes, K. S. (2017). Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment*, *161*(2), 203–211. https://doi.org/10.1007/s10549-016-4035-1

Yang, M., Kiang, M., & Shang, W. (2015). Filtering big data from social media - Building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, *54*, 230–240. https://doi.org/10.1016/j.jbi.2015.01.011

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. *NeurIPS*, 1–11. http://arxiv.org/abs/1906.08237

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, 233–242. https://doi.org/10.1145/2623330.2623715

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing*. http://arxiv.org/abs/1702.01923

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research*, *17*(6), e138. https://doi.org/10.2196/jmir.4305

Yu, J., & Qiu, L. (2019). ULW-DMM: An Effective Topic Modeling Method for Microblog Short Text. *IEEE Access*, *7*, 884–893. https://doi.org/10.1109/ACCESS.2018.2885987

Zhai, C., & Massung, S. (2016). *Text Data Management and Analysis*. https://doi.org/10.1145/2915031

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. *European Conference on Information Retrieval*, 338–349.

Zhao, Y., & George, K. (2001). Criterion functions for document clustering: Experiments and analysis (Technical Report). *Department of Computer Science, University Of*, 1–30.

Zhou, L., Zhang, D., Yang, C. C., & Wang, Y. (2018). Harnessing social media for health information management. *Electronic Commerce Research and Applications*, *27*, 139–151. https://doi.org/10.1016/j.elerap.2017.12.003

Zhu, X. J. (2005). *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, *48*(2), 379–398. https://doi.org/10.1007/s10115-015-0882-z

Zurynski, Y., McIntyre, P., Booy, R., & Elliott, E. J. (2013). Paediatric active enhanced disease surveillance: A new surveillance system for Australia. *Journal of Paediatrics and Child Health*, *49*(7), 588–594. https://doi.org/10.1111/jpc.12282

Appendix A
First stage topics keywords

Table 40 shows the first 20 keywords of each of the 14 topics of the 14 topic DMM model. Topic 13 was used for extracting VAEM - the words are mostly indications of getting a flu shot. The words are lemmatized, so the word "get" in this topic is likely to be a lemma of "got", while "be" could be a replacement of "am". Note that some topics keywords in the remaining topics are quite similar to one another, and in the more understandable 9-topic model shown in Table 41 some of these topics are amalgamated. Topic 8 was the topic that concentrated VAEM in the 9-topic model, the keywords are almost identical to those of the Topic 13 of the 14-topic model.

Table 40: First stage DMM 14 topic model keywords

1	not, do, vaccination, people, get, make, know, child, go, would, use, flu, kill, take, say, give, need, good, cancer, disease
2	vaccination, health, new, research, work, hiv, need, not, disease, flu, world, good, day, development, vaccineswork, help, public, global, great, use
3	vaccination, not, disease, child, study, autism, risk, flu, virus, new, vaccinate, infection, influenza, cause, do, use, increase, high, health, may
4	vaccination, not, health, child, autism, news, anti, cdc, new, dengvaxia, medical, doctor, parent, science, safety, say, hpv, fake, do, public
5	vaccination, outbreak, measle, health, congo, case, experimental, new, begin, campaign, news, disease, first, world, people, use, ebola, say, cholera, vaccinate
6	vax, not, new, be, get, go, love, do, good, album, see, time, make, ticket, know, look, best_stock_investment, come, vaccination, think
7	not, do, autism, people, get, vaccinate, vaccination, child, kid, anti, because, think, know, say, be, make, go, would, can, parent
8	cancer, new, research, flu, mouse, virus, could, develop, market, study, human, news, scientist, cell, use, hiv, influenza, researcher, trial, patient
9	hpv, cancer, vaccination, cervical, get, not, girl, prevent, woman, boy, year, vaccinate, health, man, do, protect, gardasil, young, age, cause
10	vaccination, get, flu, health, not, free, travel, need, child, clinic, vaccinate, hepatitis, new, measle, do, shot, shingle, protect, school, year
11	vaccination, dog, vaccinate, pet, get, rabie, cat, old, not, need, clinic, year, vet, puppy, today, home, do, come, month, animal
12	child, vaccination, vaccineswork, polio, disease, protect, vaccinate, immunization, measle, health, week, year, world, get, not, prevent, help, campaign, do, life
13	get, flu, shot, not, do, be, go, vaccination, year, sick, take, never, shoot, today, people, time, day, vaccinate, feel, need
14	flu, shot, get, not, year, season, effective, say, influenza, cdc, health, people, vaccinate, news, do, child, new, die, still, good

Table 41: First stage DMM 9 topic model keywords

1	cancer, new, vaccination, research, flu, virus, disease, health, study, develop, hiv, influenza, use, development, could, human, work, market, mouse, news
2	vaccination, dog, get, vaccinate, pet, not, clinic, need, rabie, old, cat, today, year, new, do, go, vet, free, come, puppy
3	not, do, vaccination, people, get, make, child, go, know, flu, would, use, good, cancer, need, kill, take, say, disease, give
4	flu, shot, get, not, year, season, health, influenza, vaccination, new, effective, cdc, say, vaccinate, people, child, news, shingle, do, protect
5	vaccination, health, outbreak, measle, child, polio, case, vaccinate, campaign, disease, new, congo, news, get, first, people, begin, world, year, vaccineswork
6	vaccination, hpv, child, cancer, get, vaccineswork, not, protect, health, disease, vaccinate, prevent, immunization, cervical, year, do, need, week, measle, help
7	not, do, get, people, autism, vaccinate, vaccination, child, kid, vax, anti, think, be, know, say, because, go, make, would, can
8	flu, get, shot, not, do, be, go, year, vaccination, sick, take, never, shoot, people, today, time, need, day, say, give
9	vaccination, not, autism, child, health, hpv, study, new, do, cdc, link, science, parent, say, doctor, safety, news, anti, medical, dengvaxia

Table 42 contains the keywords of a 10-topic MALLET model from Stage one. Ten topics had the highest recall of the MALLET models around this number of topics.

Table 42: First stage MALLET 10 topic model keywords

1	people, autism, kid, parent, kill, stop, doctor, fact, story, question, reason, government, mom, claim, truth, american, mandatory, dangerous, lie, vaxxed
2	develop, country, support, global, provide, program, universal, development, control, healthcare, great, population, market, influenza, vaccination, target, improve, part, research, fund
3	flu, shot, year, effective, bad, sick, time, season, shoot, cdc, doctor, influenza, die, feel, late, strain, stay, ago, give, nurse
4	vaccination, dog, free, today, call, month, pet, clinic, rabie, care, check, vaccinate, home, cat, visit, travel, animal, date, offer, love
5	vaccination, cancer, hpv, prevent, risk, school, high, patient, rate, girl, age, woman, infection, increase, adult, man, shingle, treatment, young, cervical
6	study, science, research, read, show, find, human, test, link, post, scientist, medical, safety, article, issue, trial, result, researcher, base, share
7	child, vaccinate, vaccineswork, day, protect, week, polio, world, life, baby, immunization, safe, disease, family, healthy, live, learn, hiv, community, death
8	health, vaccination, disease, news, measle, virus, outbreak, case, public, spread, hepatitis, campaign, death, state, report, fight, begin, deadly, due, epidemic
9	make, good, give, work, time, drug, hope, cure, food, big, lot, pay, body, long, money, put, problem, fake, create, inject
10	anti, vax, thing, watch, trump, back, fuck, talk, shit, video, guy, sad, play, turn, run, end, listen, head, love, fucking

Some of the topics in Table 42 have similarities to the DMM 9-topic model, but there are some that are more difficult to interpret. For instance topic 9 — it's difficult to interpret the words "make, good, give, work, time, drug, hope, cure, food, big, lot, pay, body, long, money,

put, problem, fake, create, inject" as a topic, and apart from the first words "anti, vax" and the word "trump" of topic 10 suggesting an anti-vax theme the remaining words "thing, watch, back, fuck, talk, shit, video, guy, sad, play, turn, run, end, listen, head, love" etc. do not seem to add anything to the theme.

Table 43 shows another pass of the MALLET model with a different seed value, this model was preferred to the previous one. Again, there are comprehensible topics — for example, Topic 6 seems a very good collection of words representing autism concerns — but there are others like topic 10 where it is difficult to attribute a theme. Topic 7 in both variations are almost identical but other topics are quite different between these versions of the 10 topic MALLET model.

Table 43: First stage MALLET 10 topic model keywords example 2

1	flu, shot, year, people, die, work, doctor, time, bad, sick, effective, season, shoot, feel, give, day, late, strain, miss, ago
2	vaccination, health, free, school, care, public, clinic, today, call, provide, visit, travel, offer, information, hospital, service, include, check, cost, require
3	vaccine, news, measle, outbreak, case, death, risk, cdc, report, high, rate, hepatitis, influenza, increase, adult, shingle, recommend, health, begin, due
4	vaccine, medical, make, government, medicine, work, drug, trump, universal, pay, state, create, money, problem, american, mandatory, force, job, push, fund
5	vaccine, kill, stop, live, control, food, body, find, man, human, population, inject, vaccin, safe, brain, poison, water, add, gmo, eat
6	vaccine, autism, anti, science, people, read, link, post, fact, show, safety, article, story, question, video, claim, truth, dangerous, lie, fake
7	child, vaccination, disease, vaccineswork, protect, world, polio, week, immunization, country, important, campaign, prevent, learn, safe, spread, support, part, community, meningitis
8	vaccinate, dog, today, vaccination, good, month, pet, day, love, rabie, find, girl, healthy, home, vaccinated, family, cat, animal, date, boy
9	vaccine, cancer, hpv, virus, research, study, develop, patient, prevent, woman, infection, hiv, cure, treatment, test, cervical, influenza, scientist, development, prevention
10	make, kid, people, vax, give, good, life, baby, thing, time, parent, talk, back, save, happen, put, hope, hear, fuck, mom

In contrast, the topics of the 10-topic Gensim model in Table 44 are all amenable to interpretation and are arguably the best collection of keywords per topic. Compared to the almost incomprehensible anti-vax topic 10 of the MALLET model in Table 42, the anti-vax theme in topic 1 of the Gensim model seems very coherent: "anti, vax, vaccination, make, trump, food, kill, government, right, people, smallpox, bill_gate, force, new, use, push, poison, water, pay, control".

Table 44: First stage Gensim 10 topic keywords

1	anti, vax, vaccination, make, trump, food, kill, government, right, people, smallpox, bill_gate, force, new, use, push, poison, water, pay, control
2	cancer, hpv, disease, virus, prevent, cure, cervical, vaccination, infection, could, woman, development, treatment, girl, patient, boy, use, prevention, cause, develop
3	not, do, get, know, people, go, be, think, would, say, can, want, make, good, vaccination, need, thing, really, take, work
4	child, vaccinate, disease, not, kid, protect, get, people, parent, die, vaccination, vaccineswork, life, death, year, baby, many, measle, family, risk
5	vaccination, dog, vaccinate, today, pet, need, get, rabie, clinic, free, old, congo, cat, come, date, travel, visit, year, home, check
6	flu, get, shot, year, not, shoot, sick, still, time, last, season, never, today, be, go, take, bad, arm, week, feel
7	new, effective, influenza, flu, study, news, cdc, report, shingle, universal, recommend, risk, year, vaccination, health, late, virus, adult, safe, show
8	autism, science, because, read, link, vaccination, article, doctor, question, cause, safety, video, claim, anti, medical, study, cdc, research, news, vaxxed
9	vaccination, health, outbreak, vaccineswork, immunization, measle, public, world, week, campaign, case, polio, country, begin, school, program, rate, care, dos, community
10	hiv, new, work, research, test, use, human, world, polio, day, develop, market, trial, global, first, drug, experimental, aid, create, ebola

Appendix B
Taxonomy to topic mapping

A taxonomy (Section 5.4) was derived with consideration to how the topic models segmented the data, but also incorporated distinctions made in this research, such as the differences between potential vaccine adverse event mentions and personal health mentions. Table 45 fits the percentages of posts per topic of the first stage Gensim 10-topic model (Table 44) over the taxonomy, to give some idea of the quantities of tweets per topic in the taxonomy. It is interesting to note that decidedly anti-vaccination messages are only around 8% of the tweets, and there are more tweets promoting vaccination at 9.7%. VAEM fall mainly into Topic 6, but there are some also in Topic 3, and likewise there are personal health mentions and discussions in Topic 6.

Table 45: Taxonomy to Topic mapping

Subject	Description	Topic	Keywords	%
VAEM	Vaccine Adverse Event Mentions	6	flu, get, shot, year, not, shoot, sick, still, time, last, season, never, today, be, go, take, bad, arm, week, feel	10.7%
Personal Health Mentions	Mentions of experiencing health issues but not VAEM			
Discussions	Enquiries / Discussions / Complaints mentioning vaccines — can be emotional, sensational or neutral, but not overtly pro or anti-vaccination	3	not, do, get, know, people, go, be, think, would, say, can, want, make, good, vaccination, need, thing, really, take, work	20.7%
"The Vaccines"	The indie rock group 'The Vaccines'			
Pro-Vaccination	Sentiment or language against anti-vax viewpoints, pro vaccines, including promoting and advertising vaccines, can be implicit	4	child, vaccinate, disease, not, kid, protect, get, people, parent, die, vaccination, vaccineswork, life, death, year, baby, many, measle, family, risk	9.7%
Anti-Vaccination	Obvious sentiment against vaccines — anti-vax	1	anti, vax, vaccination, make, trump, food, kill, government, right, people, smallpox, bill_gate, force, new, use, push, poison, water, pay, control	8.0%
Autism	All autism related discussions	8	autism, science, because, read, link, vaccination, article, doctor, question, cause, safety, video, claim, anti, medical, study, cdc, research, news, vaxxed	8.5%
HPV & Cancer	HPV and cancer-related vaccine discussions	2	cancer, hpv, disease, virus, prevent, cure, cervical, vaccination, infection, could, woman, development, treatment, girl, patient, boy, use, prevention, cause, develop	7.5%
Pets and Veterinary	Pet and animal related discussions, including what might be classed as VAEM had they related to human subjects	5	vaccination, dog, vaccinate, today, pet, need, get, rabie, clinic, free, old, congo, cat, come, date, travel, visit, year, home, check	10.3%

Trends and Outbreaks	Statements and headlines mentioning trends and outbreaks	9	vaccination, health, outbreak, vaccineswork, immunization, measle, public, world, week, campaign, case, polio, country, begin, school, program, rate, care, dos, community	11.7%
Research and Studies	Mentions of new studies and research, science of vaccine development, including headlines mentioning research	10	hiv, new, work, research, test, use, human, world, polio, day, develop, market, trial, global, first, drug, experimental, aid, create, ebola	6.0%
News	News articles, headlines, and announcements. Statements from vaccine-related organizations	7	new, effective, influenza, flu, study, news, cdc, report, shingle, universal, recommend, risk, year, vaccination, health, late, virus, adult, safe, show	6.8%

This is just an indicative look as the matching is not exact, and other topic models can also be fitted to the taxonomy, but as the topic numbers increase, they are more difficult to fit.

Appendix C
Second stage topic modelling comparisons

In the first stage of topic modelling, the best performing topic number (or numbers if combined) had been identified per model for each topic model architecture. These topics were referred to as VAEM topics. The best VAEM topics of the models were Topic 13 of the DMM 14-topic model, topics 3 and 6 combined of the Gensim 18-topic model and topic 3 of the MALLET 10-topic model. The texts from these were extracted as 3 datasets for testing in the second stage, where further topic modelling was performed on each dataset, using each of the three model types, testing a range from 2 to 20 topics. Nine result sets were thereby obtained and compared with one another — the goal being to see how the 3 models performed on each of the 3 input datasets, and how their results compared with one another.

Table 46 summarizes the labelled vs unlabelled documents per dataset assessed in stage two of topic modelling and shows that the datasets are much reduced from the 328,822 records of the stage one dataset. Based on the proportion of labelled VAEM vs other remaining labelled documents there is a far greater concentration of vaccine adverse event mentions in the data compared to the starting position. That is, in the first stage of topic modelling there were 222 vaccine adverse event mentions in 1,400 labelled documents - only 15.9% VAEM, but in the second stage the VAEM percentage is up to 45% of the labelled documents.

Table 46: Second stage data descriptions

Topic Model	VAEM	Total Labelled	Total Unlabelled	Total Docs	% VAEM to labelled
Gensim 18	206	530	58,244	58,774	38.9%
Mallet 10	198	488	43,679	44,167	40.6%
DMM 14	220	485	37,117	37,602	45.4%

Applying the 3 topic model types to the 3 datasets resulted in 9 sets of results. These were compared with one another. Figure 44 shows the result as a grid of the training plots of each model over each dataset, the heavier line is the Adjusted F1-Score, the dotted line is precision, the lighter solid line is recall. The best result was obtained by the DMM model over the DMM data, which is the first plot in the chart. It reached an Adjusted F-Score of 0.82 at 9 topics with both a high recall and precision.

Figure 44: Second Stage topic models vs data

The next 3 sections contain detailed analyses of each model's performance.

C.1 Gensim models

Gensim LDA models performed best with from 3 to 7 topics, with later peaks in performance around 15 topics. Generally, after 3 to 5 topics the recall was markedly less than precision. There was a lot of variability in the Adjusted F-Score, the most regular results came when using the DMM dataset, which is shown in Figure 45. The highest F-Score is around 0.72 with 3 topics due to a high recall, but the precision is very poor at 0.45. The next highest F-Score is around 0.68 at 15 topics, but at that point the recall is only 0.63. There is no point at which both recall, and precision are optimal.

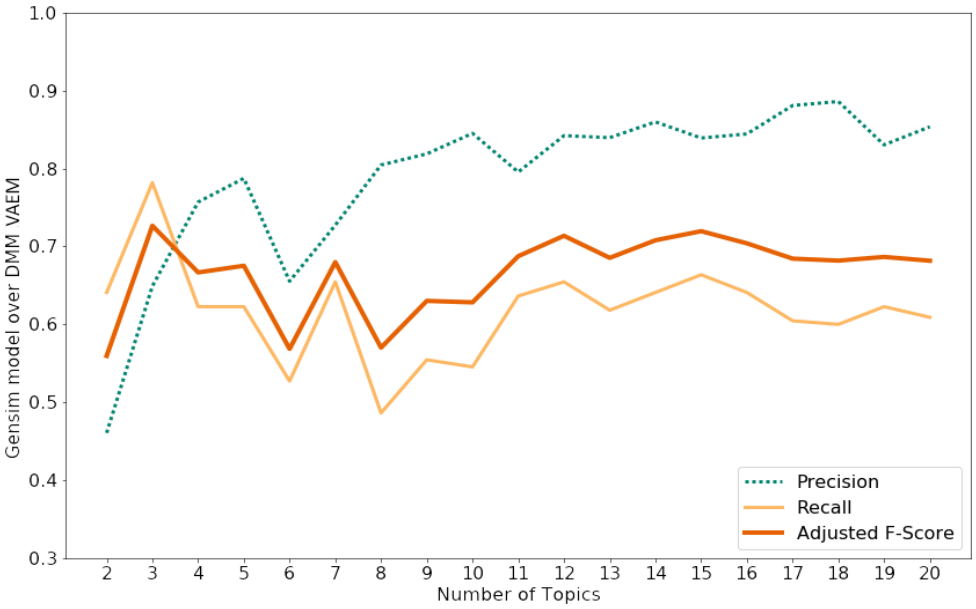


Figure 45: Gensim LDA model over DMM dataset

C.2 MALLET models and datasets

MALLET models achieved their highest recall with a low number of topics, but with an accompanying low precision. Convergence of recall and precision only occurred with high topic counts with a decrease of recall and increase of precision. This is most evident when applied to the MALLET dataset, see Figure 46. In this chart the best Adjusted F-Score by a small margin was 0.813 at 19 topics, due to a precision of 0.788 and a recall of 0.828. However, better results for recall at very slightly lower F-Scores are obtained at 18 and 17 topics.

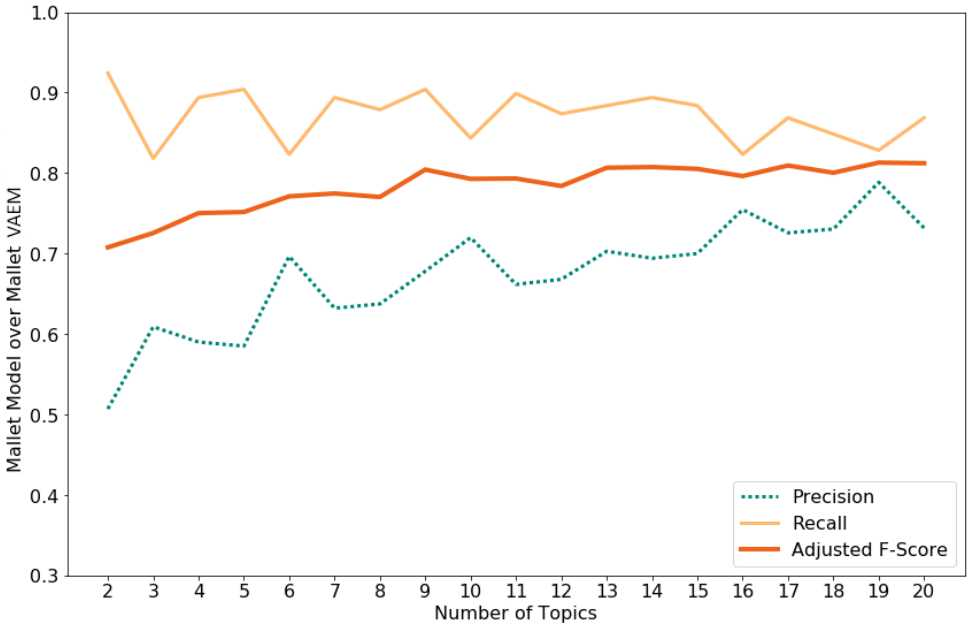


Figure 46: MALLET model over MALLET dataset

As these results were only obtained at higher topic counts, they were also discarding VAEM texts to achieve smaller grouping, for instance the 19-topic model identified in the training plot contained only 163 of the 198 VAEM that were available in the MALLET dataset.

Figure 47 depicts that the MALLET *dataset* was handled similarly by the DMM model, it also exhibited an initial high recall, a slow convergence of recall and precision, and achieved similar results to the MALLET model. This confirmed the stage one topic modelling analysis that data coming from MALLET topic models was splitting VAEM between topics, and so it was difficult to achieve an optimal balance of recall and precision in one topic. At 18 topics it had an Adjusted F-Score of 0.833, and an F-Score of 0.821, and contained 172 VAEM from the available 198.

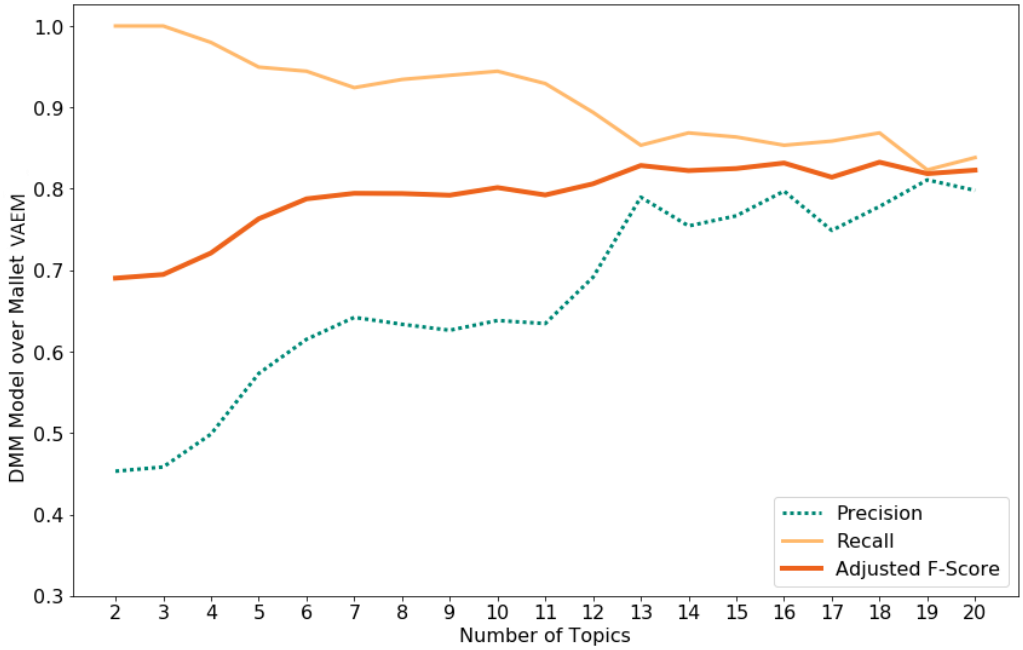


Figure 47: DMM model over MALLET dataset

C.3 DMM models and dataset

DMM models performed best in terms of balanced recall and precision from early in its training run when used on the DMM dataset. Figure 48 demonstrates the early convergence of recall and precision found by the DMM model over the DMM dataset, and that the best performing model was with 9 topics, with both an Adjusted F-Score and F-Score of 0.820 at 9 topics, due to a high precision of 0.829 and recall of 0.814. There were 179 VAEM in the best topic, out of 220 VAEM that were available in the DMM dataset.

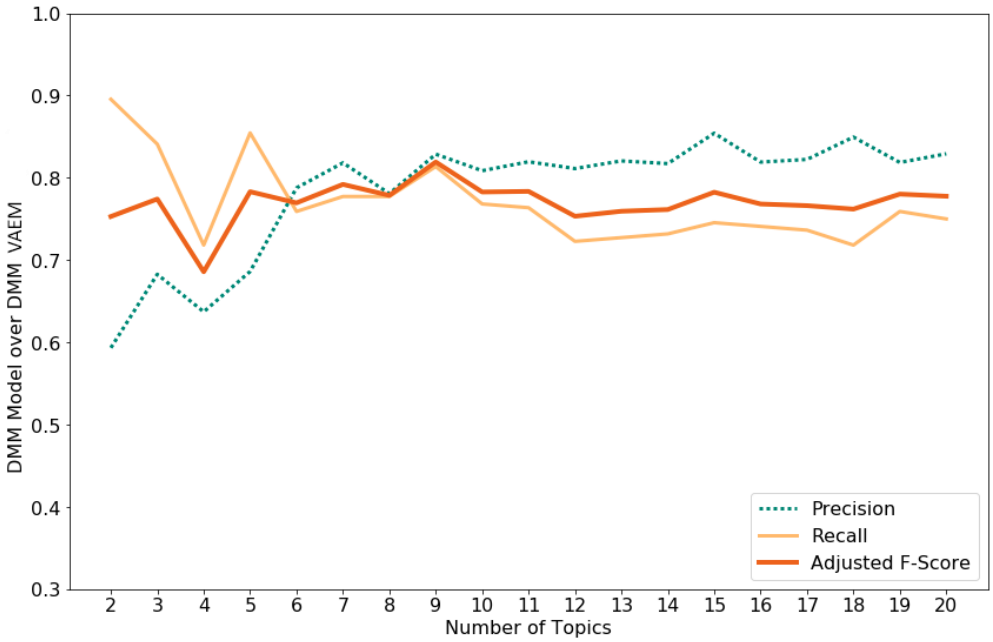


Figure 48: DMM model over DMM dataset

A contender for best combination of model and dataset was the MALLET model over the DMM dataset, depicted in Figure 49. Its best Adjusted F-Score was at 17 topics, but its Adjusted F-Score was still less than that of the DMM model at 9 topics, and with a count of 172 VAEM it identified less than the 179 from the DMM model.

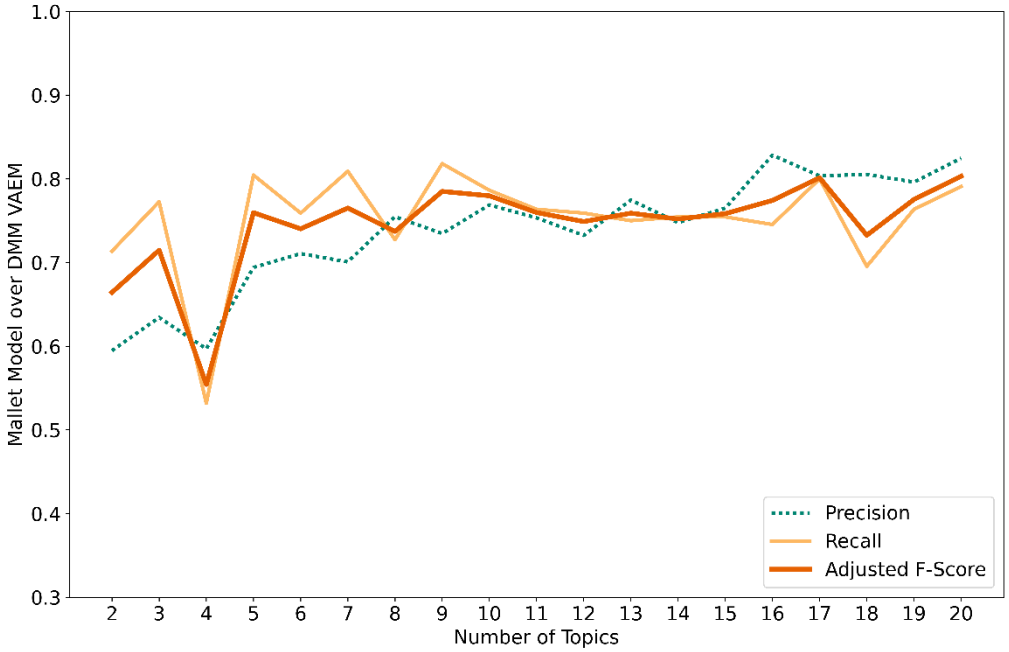


Figure 49: MALLET model over DMM dataset

Using Adjusted F-Score and recall and considering the number of VAEM that exited in the dataset the clear leader from this assessment was the DMM 9-topic model over the DMM dataset. The DMM 9-topic model over DMM data identified 179 of 220 available VAEM whereas DMM over MALLET data found only 172 of 198. The best performing combinations are presented in Table 47.

Table 47: Best second stage model & dataset combintation

Topic Model	Dataset	Topics Count	Precision	Recall	F-Score	Adjusted F-Score
DMM	Mallet	18	0.778	0.869	0.821	0.833
DMM	DMM	9	0.829	0.814	0.821	0.820
DMM	Gensim	17	0.720	0.874	0.790	0.810
Mallet	Mallet	19	0.788	0.828	0.808	0.813
Gensim	Mallet	3	0.575	0.874	0.694	0.732

Appendix D
Word embeddings and embeddings-based features

The following sections first summarize the procedures used when generating word embeddings and secondly describe how additional features were generated from embeddings to help distinguish vaccine adverse event mentions. The word embeddings were used by the Deep Learning models, and by some experiments with standard models (Appendix E). VAEM-related features derived from Word2Vec similarity scores were used by a rule-based technique, which is described in 6.10.

D.1 Vaccine-related word embeddings

A Word2Vec embeddings model was trained on the *entire* cleaned dataset. Cleaning consisted of the same approach as used in the topic modelling: converting from Unicode to plain text, then lower-case conversion; URLs, the retweet tag, the hash symbol from hashtags, and @user references were all removed. Text-based emoticons were replaced with plain English, other emoticons were removed. After some experimentation it was found that a 100-vector length discovered word relationships adequately while increasing the vector length had no discernible benefit. For comparative purposes, a Word2vec model was also trained on the lemmatized data from the topic modelling phase, which contained n-grams of commonly encountered lemmatized phrases.

The word embeddings generated from the entire collection of tweets clearly finds associations between words related to VAEM. A Word2Vec similarity function obtained the most similar words around some of the key indicators - the words “pain”, “arm”, and “sore”, the top 10 results are tabulated below in Table 48 and Table 49. Depending on whether the source text is the original cleaned text or the lemmatized version slightly different results are obtained, but it is evident that “arm” refers to “sore” and “hurts” in the non-lemmatized version of the similarity table and that “arm” and “sore” are both associated with the bigram “arm_hurt” in the lemmatized version of the table. The Word2Vec model trained on the cleaned text was used with the neural networks and with the rule-based classifier.

Table 48: Top 10 similar VAEM words

pain	arm	sore
soreness	aching	aching
redness	hurts	ache
rashes	arms	bruised
headaches	sore	achy
pains	bruised	headache
fatigue	swollen	itchy
muscle	throbbing	achey
swelling	shoulder	aches
nausea	bruise	runny
fainting	ow	congestion

Table 49: Top 10 similar VAEM n-grams

pain	arm	sore
headache	sore	ache
soreness	left	arm_hurt
nausea	arm_hurt	bruise
anxiety	bruise	achy
ache	hurt	feverish
redness	ache	throb
unbearable	throb	swollen
fatigue	upper_arm	blood_drawn
relieve	swollen	soreness
swelling	arm_feel	body_ach

D.2 Word embedding cluster features

Using unsupervised word representations as additional word features has been an effective way to improve accuracy in NLP tasks (Nikfarjam et al., 2015). To evaluate this, an experiment was conducted to determine whether Word2Vec word embeddings could be used as an alternative to, or additionally with, the count-based vectors traditionally used with traditional classifiers — see Sections E.1 and E.2 of Appendix E.

After training Word2Vec word embeddings on the entire unlabelled corpus, K-means clustering was performed on the word embeddings. K-mean clustering is a popular, simple approach for grouping similar words (Arora & Varshney, 2016). Alike words were clustered in a group based on their place in vector space. As well as providing alternative vectors for classifiers this assisted with identification of top VAEM words for the rule-based approach.

D.3 Clusters for a rule-based approach

K-means clustering was performed on the entire data and after experimentation with different cluster numbers it was decided that the optimal number for a rule-based manual approach was 150 clusters. 150 was a number that could be handled in a manual exploration of the data, see examples in Table 50. The first two examples in the table are both from cluster 80, but the first is in order of the words found most frequently in VAEM posts, the second entry is in order of word frequencies in non-VAEM posts. The most important cluster 80 words ("arm, hurts, arms, hurt, feels, shoulder, hurting") in the VAEM-related posts indicate the effects of a recent vaccination.

Table 50: Cluster examples using 150 clusters

Cluster	Words
80 in VAEM order	arm, hurts, arms, hurt, feels, shoulder, hurting, lift, punched, ouch, needle, bruised, drawn, shoulders, needles, ow, throbbing, butt, bruise, stabbed, deltoid, swelled, jabbed, workout, lifting, fainted, cried, bicep, lollipop, tattoos, faint, cheek, bruises, tattoo, weights, wrist, amputated, champ, noodle, piercings, itches, bandaids, poke, thighs, poked, pinch, pokes, lightheaded, screamed, thigh, ankle, piercing, knot, wimp, stab, bandaid, bled, wuss, finger, stitches, pierced, poking, cushion, prick, flinch, pricked, bandage, tequila, sucker, bleed, phobia
80 in Non-VAEM order	needles, needle, tattoo, tattoos, piercings, stabbed, drawn, bandaid, stab, prick, poked, piercing, tequila, phobia, cheek, cried, poke, finger, bleed, bandage, flinch, cushion, pierced, jabbed, wrist, poking, ankle, sucker, pricked, wimp, lollipop, stitches, noodle, thigh, bandaids, workout, pinch, itches, screamed, thighs, bruise, lightheaded, bled, bruises, pokes, wuss, knot, weights, amputated, champ, lifting, bicep, faint, butt, deltoid, fainted, swelled, ouch, shoulders, ow, bruised, throbbing, punched, lift, shoulder, hurting, hurt, feels, arms, hurts, arm
51	my, got, feel, yesterday, sleep, sick, hours, im, bed, tired, night, bit, woke, felt, ugh, barely, forgot, hour, ill, minutes, gotten, kicking, sleeping, exhausted, sucks, hella, asleep, nap, slightly, glad, cranky, ive, soooo, fml, remembered, sooo, lowkey, asf, knocked, crappy, stuck, napping, throwing, drowsy, sooooo, naps, ached, hives, dunno, couch, sitting, owie, laying, realized, deathly, hurty, groggy, pounding, mins, weenie, grumpy, yay, miserable, forgetting, ish, puked, bleh, ughhh, darn, whining, knock, woozy, stings, shaking, stressed, lethargic, oof, wakes, dreading, puke, bragging, awhile, terrified, feelin, ughh, regretting, welp, invincible, hrs, resting, teething, alllll, dread, shivering, havent, anxious, mofo, crummy, afterward, skipped, ducking, bih, er, hubby, yday, meh, yucky, sweating, stayed, mend, kiddo, mildly, friggin, shoulda, advil, sucked, def, pinched, congested, kms, depo, sniffles, relieved
146	pain, reaction, effects, painful, symptoms, allergic, swelling, rash, anxiety, itching, migraine, ibuprofen, mild, fatigue, lymph, cramps, flare, strep, tenderness, exhaustion, intense, slight, minor, pains, vomiting, stiffness, fevers, stress, temporary, steroid, appetite, migraines, lethargy, inflamed, redness, nodes, excruciating, unwell, fainting, reactions, rashes, tamiflu, distress, tylenol, discomfort, bleeding, lingering, traumatic
279	sore, feeling, throat, muscle, swollen, aches, numb, headache, aching, slept, ache, chills, soreness, dizzy, runny, sinus, muscles, achy, nauseous, stomach, sleepy, joints, stiff, headaches, nausea, congestion, feverish, itchy, achey, stuffy, sweats
313	fever, yellow, yellowfever, brazil

The importance of needles, tattoos, and piercings in non-VAEM posts reflects many posts that indicate an ironic aversion to vaccine injections, when the subject has extensive tattoos

and piercings and therefore ought not to be frightened. References to bandaids and lollipops etc. are also indicative of a recent vaccination, but there are twelve times as many mentions of bandaid in the non-VAEM group, the conversations are typically saying that the vaccination was uneventful but that a decorative bandaid was a good outcome!

D.4 Clusters for alternative vectors

When using the clustering approach for classification by applying the centroid of words as a feature (Section E.2), it was found that experimentation with cluster size was more useful than assessing numbers of clusters. A smaller size of around 20 produced a good outcome — which resulted in 3,170 clusters, examples in Table 51. It was observed that highly correlated words clustered together — for instance "needle, phobia" in cluster 1573 and "needles, piercings, tattoos" in cluster 2172 — they indicate similar but distinct ideas, and both are heavily used in non-VAEM rather than in VAEM. On the other hand, some of the useful groups that were found with the larger clusters used for the rule-based approach were split, cluster 1295 has "fever, yellow" but we need to go to cluster 1254 for "brazil, yellowfever". The larger cluster 313 in Table 50 had placed all of these together, which was reflected in the texts — many posts with yellow fever vaccination mentions were in the context of travel to Brazil.

Table 51: Cluster examples using cluster size 20

Cluster	Words
2867	arm, shoulder, sore
2070	ache, aching, bruise, hurting, hurts, numb, ow, punched, stiff, throbbing
736	bleed, excruciating, itch, itching, minor, nerves, pain, painful, scars, scratch, sores, uncomfortable
761	adverse, effects, reaction, reactions
1295	fever, yellow
1254	brazil, yellowfever
1573	needle, phobia
2172	needles, piercings, tattoos
1987	got, sick, gotten, hella, sucks, everytime, regretting, caught, feelin, sucker, killin, faint, puked, puking, shoulda, allllll, bedridden, blizzcon, bouta, clinicals, coughed, crud, deathly, def, dodging, hypochondriac, invincible, ish, knocks, probs, shouldve, skipped, sneezed, sniffle, stg, streak, tequila, twitchcon, welp, yeet

Appendix E
Assessment of embeddings as vectors and features

The features machine learning models learn from are the numeric vectors that represent the data they are processing, which for NLP tasks are the sparse word vectors or dense word embeddings created during pre-processing. Additional features can also be engineered from the data to help guide the model to understand its data more clearly. Experiments were conducted using embeddings as alternative to the standard sparse vectors used with traditional classifiers, and additionally as added features. The experiments used an Extra Trees classifier, which due to its architecture is an effective traditional classifier for conducting this kind of experiment, and which had performed well with the standard approach.

E.1 Word2Vec embeddings as vectors

In the second phase of evaluation, two experiments were conducted using word2vec embeddings as an alternative source of vectors for the traditional classifiers. That is, instead of creating sparse matrices using TF or TF-IDF vectors of one-hot encoded words, the words' word2vec embeddings were utilized.

One approach was to take the average of the word2vec embeddings of all words in a document, which potentially has the effect of locating the entire document into a vector space, with the result that similar documents have similar scores.

The other approach was to use K-means clustering to create clusters of words based on their similarity, and then use the K-means centroid as a cluster number to assign to the words in the document. Experimentation with cluster size showed that dividing word embeddings by 20 produced a reasonable outcome — which resulted in 3,170 clusters from the 63,403 words available, with cluster sizes ranging from 1 to 1,839. If a larger size of 200 was used to get 317 clusters, the maximum cluster size was 2,599 — but these did not perform nearly as well, perhaps too many not very similar words were assigned the same cluster number.

The results of these experiments showed that there was no advantage over the standard approach. For instance, the Random Forest classifier trained on the final dataset using the standard approach achieved an F1-Score of 0.8249 on the validation dataset; but trained with Word2Vec embeddings, it was 0.8180, and with Word2Vec centroids from the same dataset the F1-score was only 0.8035.

However, embeddings were essential for achieving the best results with the neural networks — here they came into effect as a much better initialisation of the words than random starting values.

Additional features were based on the words obtained from the baseline rule-based technique (Section 6.10), which identified the most significant VAEM-related and non-VAEM-related words, measured by their Word2Vec similarities to target words. The targets words were decided by clustering (Section D.2).

The features were:

- the word2vec embeddings of the words (Section E.2)
- the word2vec similarity scores to the target words (Section E.3)
- TF-IDF vectors created from the words (Section E.4)

Features were all converted to vectors, either the 100-long Word2Vec embeddings, TF vectors, or simpler *numpy* numeric arrays. Python *hstack* or the SKLearn FeatureUnion library were used to add features to existing vectors.

E.2 Word2vec embeddings as an additional feature

This experiment used the 100-long Word2Vec embeddings of the top *n* VAEM-related words (Appendix D and Section 6.10), from the top 5 down to just the individual top word. The embeddings were combined with the existing TF-IDF vectors to function as added features. The results were poor in all experiments, compared with just using sparse vectors, or even with the embeddings approach of Section E.1.

To provide an additional comparison between adding top *n* word embeddings to the TF-IDF vectors, a test was made using just the top *n* word embeddings as the only feature. The benchmark F1-Score over validation data of the standard TF-IDF sparse vectors approach was 0.8249, the best F1-Score for using top *n* word embeddings alone was for two words at 0.7590, and when adding embeddings as an additional feature to TF-IDF the best F1-Score was 0.7634.

Next, the non-VAEM embeddings were added and tested in the same way, and it was noted that while the scores improved, they were still significantly worse than the scores obtained by TF-IDF sparse vectors alone. In summary, using word2vec embeddings of the top words as features had a markedly negative effect and seemed to be just adding noise.

E.3 Word2vec similarity scores as a feature

Similarity scores were determined during the rule-based classification experiment (Section 6.10) but using the scores as a feature did not improve the results. The best F1-Score when

these were used as the only feature was 0.7126, and when combined as an additional feature to TF-IDF the score also decreased — being 0.8227 compared to the benchmark score of 0.8249. The conclusion was that using similarity scores also added nothing but noise.

E.4 Term Frequency (TF) vectors of top similar words as a feature

To evaluate the top similar words themselves, rather than their dense embeddings or scores, standard term frequency (TF) vectors of the top words were created per document, ranging from the top 5 down to the single top word. These vectors were then combined with the standard TF or TF-IDF vectors of the documents, which meant that a document was then represented by two sparse matrices. This approach yielded positive results, with the best outcome obtained from using a TF vector of the single top VAEM word combined with the standard document sparse vectors. An F1-score of 0.8360 was obtained with this approach, compared with the standard score of 0.8249. An additional experiment adding the single best similarity score to this combination, an increased recall but poorer precision resulted in a worse F1-score.

Appendix F
Feature engineering results

Table 52 presents the F1-Scores calculated over validation data for each of the experiments. The F1-score from a default implementation of the Extra Trees classifier trained on TF vectors is first supplied as a benchmark, then for each experiment two values are shown — first the F1-score, then below it the difference from the default F1-score. Note that after tuning the Extra Trees classifier the validation F1-score increased, but the benchmark score used here was obtained using the default classifier settings.

Table 52: Features experimentation scores

	VAEM-related feature	**Description**	**Best Score & *Difference***
	Standard TF vectors	Standard sparse vectors over all the document words	**0.8249**
1	Top Word	Term frequency (TF) vector of the single top VAEM-related word as an added feature	**0.8360** ***0.0111***
2	Word2vec embeddings of all text	The average of word2vec embedding vectors for the entire document, as the only feature	0.8180 *-0.0069*
3	Word2vec centroids of all text	The word2vec centroids for the entire document, as the only feature	0.8035 *-0.0214*
4	Word2vec embeddings top words only	Word2vec embedding vectors for each of the top two VAEM-related words, as the only feature	0.7590 *-0.0659*
5	Word2vec embeddings top words added	Word2vec embedding vectors for each of the top two VAEM-related words, as an added feature	0.7634 *-0.0615*
6	Similarity scores only	Similarity scores of the top five VAEM-related words, as the only feature	0.7126 *-0.1123*
7	Similarity scores added	Similarity scores of the top five VAEM-related words, as an added feature	0.8227 *-0.0022*

The best experimental result (Section E.4) is shown in row 1 of the table - it added a TF vector of the single top VAEM-related word to the underlying TF vectors of the model and improved the F1-Score by 0.01. Experiments that used either the average of Word2Vec scores of the entire text, or an array of Word2Vec centroids (Section E.1) are included in rows 2 and 3. They were a complete alternative to using the standard TF-IDF vectors and performed better than the remaining experiments, but on the validation data their F1-Score was 0.01 worse than

the standard approach, though the centroids approach did remarkedly well on the Victorian test data in Phase One of the classification test, see Figure 36.

The remaining tests were varying degrees worse than the standard approach, and contributed nothing useful. Not listed in the table, but also considered, was to add the topic model topics as features. However, this did not make sense, as all the data used in classification came from only one class of the first-stage topic model, and using the topics of the second-stage topic model as an overlay would only have been justified if the goal was a multi-class classification that aligned with those topics. Instead, as the goal was to score only one class of a binary classification, adding features that suggested other distinctions in the data was not desirable. However, an assessment of using the second-stage topics was made, and the single additional feature made no difference to the model.

It was concluded that adding engineered features was not fruitful due to the work required for any marginal positive effect, when more reliable and reproducible techniques were available by tuning the model or by choosing more powerful classifiers. The same conclusion was reached with experiments where the TF or TF-IDF sparse vectors were replaced with Word2Vec embeddings — the standard approach was more straightforward, reliable, and most often resulted in more accurate classifiers — with the promise of even better results using Deep Learning approaches. However, it was a worthwhile experiment as it confirmed that readily available classifiers and data preparation were entirely suitable for the task, but that an empirical assessment of feature engineering had been performed, for the benefit of fellow researchers who might consider similar approaches.

Appendix G
Textual analyses of errors and VAEM per topic

G.1 Classification Errors analysis

This section makes observations about why some records may have been misclassified. Table 53 shows records the RoBERTa Large model misclassified in the Victorian test dataset.

Table 53: Misclassified Victorian test data

Misclassified as not VAEM (false negatives)

1	the meningitis vaccine make my arm feel like it just got punched repeatedly
2	my flu shot arm is throbbbbing
3	I left deltoid is either A. extremely sore from yesterday's workout or B. from my flu shot today. i'm going with B
4	trying to sleep on your flu shot arm- no bueno
5	"Finally got the flu shot. My arm is killing me and I am dead tired! Would go curl up on the couch were it not for this ""work"" thing."
6	Yo the flu shot made me sick wtf
7	Why the hell did I get a flu shot the day before the PFA Rip my arm
8	I got a vaccine today and my arm is numb hey sisters
9	Go to doctors to get a lump checked. Come out with sore arm from flu shot! Lump isn't worrisome.
10	I dont understand how the flu shot makes my arm so sore

Misclassified as VAEM (false positives)

11	YAY!!! My shot was done today by an intern, and it still really didn't hurt!
12	This is the 3rd shot of the vaccine, every time it feels bad in the same way, as if I over worked out lifting. Considering it is the last shot of the series, so it's okay.
13	SINC CHILREN MERCY FORCE MI TOO TAK THEI FLU SHOT EAT 6 DES MORNING I BECME BLUND EN ONE EYE AND IM LOOSING FEELING ON MY HOELE LEIFT SIEDE
14	I would get sick after getting the flu shot
15	Just got back from getting my flu shot. I wonder if this one is going to make my arm sore or not. Some years seem to be better than others for that.

False Negatives:

It looks like some of the missed VAEM might be attributed to either unrecognized or infrequently found words, or words that are normally attributed to the non-VAEM label. For instance, the word “throbbbbing” in row two does not appear in any training data. “no bueno” in row four appears in only 6 records in the training data and two of them are non-VAEM records. “make” in row one appears more slightly often in non-VAEM in the training data, so it is possible that the statistical possibility of this being a non-VAEM influenced the decision

here. However, “arm” and “punched” which are also in row one, are found together in 95 VAEM vs 5 non-VAEM records in the training data. Similarly, the words “arm” and “sleep” found in row four appear together in 139 VAEM records vs only 21 non-VAEM records in the training data, and “arm” and “numb” seen together in row eight are found together in 61 VAEM records vs 12 non-VAEM records. It is unclear why the classifier got these wrong.

Some of the other combinations are more ambiguous, “flu shot” appears together with “sick” in the training data only 1.4 times more often in the VAEM records. “rip” appears only 1.3 times more in VAEM records, though when combined with “arm” the ratio is 3.8.

Finally, there are some texts that are quite subtle in their attribution of VAEM, rows three and nine for instance, where there is information that supports a reaction combined with other information that does not.

False Positives:

Here things are a little clearer, rows twelve to fourteen could all be correctly labelled as VAEM, row thirteen possibly wasn’t labelled as VAEM because of the exaggerated language and strange spelling but it has the structure of a VAEM, and twelve and fourteen could have been rightfully labelled as VAEM. Row eleven has a negation without which it would have been an adverse event, and row fifteen has a future tense and a question of wondering if a sore arm is immanent — these are subtle clues that are easy for a human to detect but not so for the classifier. It was observed that negations were not often found in VAEM, and as there was no specific negation-handling added, then examples like this are not surprising. However, there are very few of them, especially from the more powerful classifiers.

Table 54 contains examples of misclassified records from the larger test dataset, again from the RoBERTa large model. The same patterns are observed: false positives and negatives might result when the classifier was unable to decipher subtle language clues such as negations (e.g., example 18: “Got a sore arm...feels like I got the flu shot in it. I didn't” and example 1: “"You won't get sick from getting a flu shot" ...liars. Because guess what?? I'm sick”) and subjective judgments only a human labeller could make (e.g., the non-VAEM in example 17: “Got a flu shot today.... 10 minutes later I sneezed. Goodbye world”).

In other cases, ambiguity may contribute to incorrectly detecting the label. In example 3 the side effects that are listed are typical of flu rather than of flu shot: “I'm sick. I have mild fever, muscle aches, lock jaw, sinus is crammed full, and sore throat…”, it’s only the preceding rather subtle attribution by the post’s author that demands a label of VAEM: “10/10 do not recommend getting flu shot. Pretty sure that's why I'm sick”.

There are cases where the classifier has correctly identified records which look to have been incorrectly labelled: examples 7 to 12 and examples 19 to 24. They have left these intact for this analysis, but any incorrectly labelled records need fixing in the published dataset.

Table 54: Misclassified larger test data

Misclassified as not VAEM (false negatives)

1	"""You won't get sick from getting a flu shot"" ...liars. Because guess what?? I'm sick"
2	"That sadistic whore who gave me my flu shot yesterday was like ""Your arm won't hurt afterward."" Lying bitch."
3	10/10 do not recommend getting flu shot. Pretty sure that's why I'm sick. I have mild fever, muscle aches, lock jaw, sinus is crammed full, and sore throat. Don't do it. It's not worth it.
4	Got the flu shot then got a cold 3 days later.. my immune system ain't shit
5	bro i can feel myself getting sick, and I just got the flu shot at work, probably wasn't the smart idea lmao.
6	Feels like there is something tickling the inside of my ear and the injection site from my flu shot is lumpy and hurts.

Correctly classified as not VAEM, incorrectly labelled as VAEM

7	"As I sat down to get my flu shot my amazing sister came over to hold my hand bc she knows shots have made me nervous since I was a kid. As soon as the needle was in my arm I hear ""Holy crap that needle is big"". Thanks @DizzityDanielle, you're the best."
8	It's moments like when Anthony kisses my shoulder where I got my flu shot to make it feel better where I'm just like... swoon #thewarmfuzziesandshit
9	Boss : flu shot in the neck or the arm?..... Me: Moddafucken uhhhh
10	Got my flu shot and PPD today and had to take my top off so the nurse could get to my arm. It's the most action I've had in months. FML
11	Ooh, this beer makes me feel much better- screw you, flu shot. 10 min later: uh oh
12	Headed for my flu shot, RIP my arm

Misclassified as VAEM (false positives)

13	Exactly! These vaccines will have you out on sick leave
14	Feeling sore and worn out from my kickboxing and flu shot yesterday. Who wants to make me breakfast? I'll take a greek omelette and some toast, please...
15	Flu shot this morning at work. I'm left handed and my left shoulder already gets sore but it's easier to sleep on my right side so should I get it in my left arm?
16	Fuck a flu shot these side effects are
17	Got a flu shot today.... 10 minutes later I sneezed. Goodbye world
18	Got a sore arm...feels like I got the flu shot in it. I didn't

Correctly classified as VAEM, incorrectly labelled as not VAEM

19	got a flu shot yesterday and now i feel like i got the flu
20	I woke up twice now because I roll over on where I got my flu shot today.
21	I would get sick after getting the flu shot
22	Just had meningitis vaccinations and it wasn't even that bad until now lmao aches like hell <sad>
23	Of course my throat is itchy day after my flu shot
24	This is the 3rd shot of the vaccine, every time it feels bad in the same way, as if I over worked out lifting. Considering it is the last shot of the series, so it's okay.

G.2 Vaccine adverse event mention examples per topic

Table 55 contains examples of vaccine adverse event mentions, per topic of the second stage DMM 9-topic model, taken from the data collected in the second phase of classification. There is not much to distinguish them, but the majority of the VAEM (6,320) are found in Topic 8, and it has many tweets like the examples here that focus on having painful arms after getting a flu shot. Other topics contain fewer potential VAEM. For example Topic 5 contained only 6, the first 4 show here are very similar to one another: "this/the flu shot…".

Table 55: Vaccine adverse event mention examples

Topic 1 – 102 tweets
My son's has a fever and his arm aches from vaccinations yesterday. The only thing I can think of that would be worse is polio, measles, smallpox, etc. Vaccinate your kids! #VaccinesWork #vaccinessavelives
that meningitis vaccine hit me real bad oof.....
This is my poor baby 2nd baths because she's been running a fever since yesterday because of the vaccines
Who cares not everyone needs to nor can be vaccinated, I got vaccinated once and my face and my arms swelled up and I got rashes on my arm I was pissed
Topic 2 – 70 tweets
that flu shot fucked me up
I got my flu shot and I feel like I'm dying
I got a flu shot and my arm hurts so bad and my mom basically called me a punk.
Flu shot got my right delt looking swole rn
Topic 3 – 22 tweets
can't focus, flu shot kicked my ass, time to die
Flu shot swelled my arm
Me: ouch my arm hurts from my flu shot Mother: drink another glass of wine or two and you'll forget about it She's so wise
Yesterday I discovered that the flu shot is a migraine trigger for me. So that has sucked.
Topic 4 – 244 tweets
my arm is so bruised from my flu shot and i don't think that's normal is this a sign that my time has finally ARRIVED...?
My arm hurts never getting a flu shot again by anna
Did anyone else get super sick from the flu shot???!?!!?
Who gets a flu shot when they are already sick? This guy!! Who gets sicker?! This guy!!
Topic 5 – 6 tweets
This flu shot got me hella sick
This flu shot hittin different
The flu shot made me get sick and nose is stuffy yet runny at the same time like???
The flu shot was not worth it I am so sick haha

Topic 6 – 139 tweets

This flu shot got me even more sick
This flu shot got me fcckkkddd up. Don't do it ppl
Only I would get the flu shot and have an allergic reaction
This years flu shot had me in the hospital at 1am w allergic reaction, don't get that shit fam

Topic 7 – 720 tweets

I get the flu shot and right after I get sick. Awesome
That flu shot tore my are all the way up.
Forced to get the flu shot now I feel terrible
I got my flu shot and no one told me I would get sick until after.... when I got sick. Fuck y'all.

Topic 8 – 6,320 tweets

Got my flu shot and now my arm sore af
Just got vaccinated.. My arm is numb right now
got a flu shot yesterday and my arm swollen asf
I got a flu shot yesterday and my arms so fucking sore bruh

Topic 9 – 534 tweets

Turns out getting a flu vaccine gives you the symptoms of the flu Huge scam
My arm is just starting to recover from the flu shot. My ass was disabled for a few days.
Flu shot made me feel miserable for about 12 hours. What a ride.
My work made me get a flu shot.. its immediate soreness.

Appendix H
Victorian data analysis

As described in Section 4.6.3, for investigating the applicability of the data to local seasonal safety-signal trends, data coming from Victoria, Australia was identified by using the geographical-related Twitter fields UserLocation, UTCoffset, TimeZone and Place. Victoria was chosen as that is the author's local area and the trends identified in the data could be checked with the local vaccine safety reporting authority AEFI-CAN. The data covered the period between 7th February and 7th June 2018, which includes the time when people were getting flu vaccines, and the early flu season of 2018. The numbers were however relatively low, 3,112 tweets were found and labelled, with 90 vaccine adverse event mentions records and 3,027 non VAEM.

The weekly incidence of discussions is plotted in Figure 50. The chart shows that the pattern in VAEM-related tweeting follows a distinct trend of increasing over the month of April and then tailing off again after mid-May. This is in accordance with the rate of flu vaccinations in Victoria, and it is a somewhat different pattern to the more general vaccine-related discussions found in the non-VAEM related texts — these do not exhibit such extreme differences and have an additional peak before the flu season. The datasets have been assigned different y-axes in the chart so the smaller numbers of the VAEM data can be clearly visualized. AEFI-CAN confirmed that these trends follow those of their surveillance reporting.

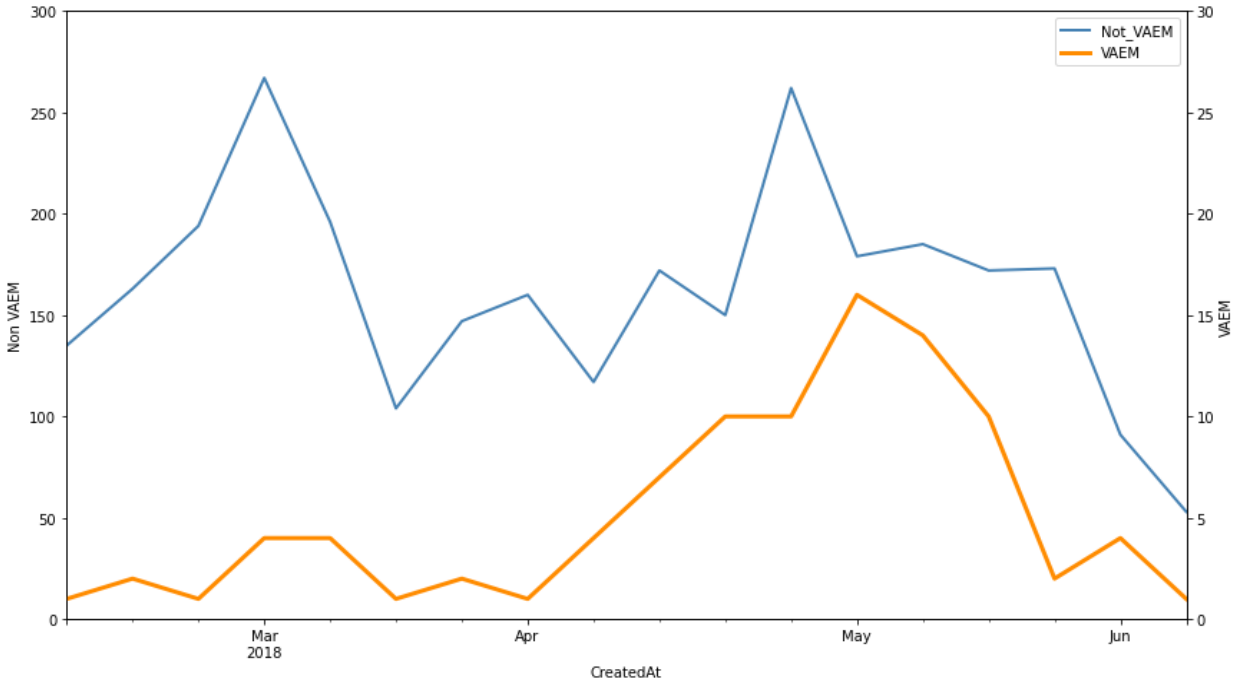


Figure 50: Victorian vaccine-related tweet trends

Appendix I
Reddit data analysis

One of the data assessments for the research used Reddit data that had been made available on the internet. This comprised of extremely large JSON files containing all Reddit submissions in one data set, and the related comments in separate data sets. The first publication of this data consists of partial data from 2006 and 2007, and all data from January 1st, 2008 through to August 31st, 2015; and thereafter Reddit submissions and comments are published monthly. Much of the data is also available to Google BigQuery (https://cloud.google.com/bigquery), where aggregate queries are freely able to be created in a sandbox, though limits are soon reached and after that query data must be purchased.

The Reddit submissions file (*Full Reddit Submission Corpus*, 2015) is a 263 GB JSON file (downloadable as a 43 GB compressed file) containing around 200 million submissions. These are the initial posting of a Reddit conversation, and may only contain a *Title* with text of interest, but sometimes will also contain the beginnings of a conversation — in a field called *Selftext*. The entire conversation initiated by a submission consists of many comments, and can be found in separately published comments archives (*Reddit's publicly available comment dataset*, 2015). The equivalent comments archive for the period contains around 1.7 billion JSON records and is over 1 terabyte uncompressed.

The Reddit data analysis used the Reddit submissions, which were imported into a SQL Server database. During importation, the JSON files were processed into individual records containing the most useful fields (Date, ID fields, the Subreddit, and the Title and Selftext fields), and the JSON itself was also retained in a field in case it might be required for later processing. There were 196,531,736 Submission records obtained.

To extract vaccine-related Reddit submissions from the data a series of queries were developed that searched for specific words, "vaccine", "vaccination" and "vax", "flu" and "influenza", but also "virus" and "viral". Text needed to additionally contain commonly used terms associated with either reactions to vaccines (e.g., "sick", "tired", "ill", "sore", "reaction"); or with descriptions of specific vaccines (e.g., "DT", "TDap", "MMR", "pentacel"); or vaccine and viral related illnesses (e.g., "measles", "mumps", "rubella", "meningococcal", "pneumococcal"); or with discussions about vaccines (e.g., "mercury", "thermisol", "adjuvant"). It was found that medical terms were very infrequently used and that widening the search to consider possible variations in spelling or colloquial equivalents introduced more noise than value. The queries also needed to incorporate a great many filters

to remove irrelevant discussions — especially because many of the terms employed are commonly used in other contexts, for instance “virus” is heavily used in computing related discussions. Useful submissions tended to be found in a group of subreddits concerned with obtaining advice - such as “doctors”, “medicine”, “medical”, “health”, “NoStupidQuestions”, “AskReddit” and “askscience”.

Table 56 presents a sample of extracted VAEM-related Reddit posts. The posts reflect the type of conversations found in Reddit — they were longer and more detailed than the VAEM-related posts that were found in Twitter, and these samples are on the brief side. When they did contain adverse event mentions then they tended to mention more serious reactions. But they were also very much less frequent and did not exhibit the direct, immediate, and casual nature of tweets — which had made tweets so valuable as a possible source for noting emerging trends.

Table 56: Sample of VAEM-related Reddit posts

My dad just 13 days ago received his flu shot from this brand Novartis. 4 days ago he came home feely extremely dizzy as if were drunk. he slept and next morning had no control on left side of body (face and legs and arm) and till this day is unable to use his left arm and limps his leg around. he drools and slurrs speech and says has trouble swallowing foods. Could it be that this flu vaccine has these side affects? i googled some stuff and even read about GBS? Guillain-Barré syndrome. Anyone please!
Took the last Gardasil vaccine yesterday. Now have high fever, body aches and chest pain. Anyone else get this series of shots and have a similar reaction?
So I got the HPV vaccine at the doctor today and about two minutes after I fainted and had a seizure spell. After the doctors got me stable again I read the page about the vaccine and it said this at the bottom: "Brief fainting spells and related symptoms (such as jerking movements) can happen after any medical procedure, including vaccination. Sitting or lying down for about 15 minutes after a vaccination can help prevent fainting and injuries caused by falls. Tell your doctor if the patient feels dizzy or light-headed, or has vision changes or ringing in the ears." I feel fine now and I'm just curious as to why this happened. If anyone knows then I would really appreciate an answer.
Late reaction to vaccinations? I received some vaccinations last week on the 17th which were: - yellowfever - Hep A/ Typhoid (vivaxim) - Fluvax along with an oral vaccination of Cholera and since yesterday (24th) I've been having bad pains in my head generally when i move my head or eyes, could this be a late reaction to my vaccinations or should i go and get it checked? I was going to book an appointment if it continued tomorrow at work.

After filtering there were around 360,000 records but searching for discussions about vaccines from within the possibly vaccine-related subreddits (i.e., medical, parental, vaccine etc. subreddits) produced only 44,600 candidate VAEM records.

After manually labelling nine thousand posts it was estimated that around 0.6% of the filtered potentially useful data was suitable for examples of VAEM, which would be only around 250 records — that is from 7 years of data! Although the records were useful, there were not enough for evaluating any machine learning-based techniques, and not enough for

this study. By contrast, Twitter data was entirely suitable for measuring direct cause and effect relationships and emerging trends in vaccine adverse event mentions.

To give some idea of the ratios of various terms for a flu vaccination in the 44,600-record set, a search for “flu shot” produced 507 records, 27 of which were a VAEM, and a search for “flu vaccination” found 362 records, of which only 4 were a VAEM. Rarely used terms included “influenza vaccination”, with 49 records; “influenza injection” with 25 records, “influenza shot” with 3, and “flu jab” with 9 records. None of the posts that were found using the rarely used terms were a VAEM. Further analysis was based on searching for specific additional words — for instance 634 records were obtained with a mention of the word “reaction”, and 30 of these contained a VAEM.

An analysis of the frequency of the term “MMR” in Reddit data is instructive: a BigQuery search was conducted over all Reddit submissions between 2015 and August 2019 and looked for submissions counts per subreddit with a mention of “MMR”. The total submissions count over the top 1,000 subreddits was 174,872, with 161,132 posts in the top 100 subreddits. Almost all the top 100 subreddits were devoted to gaming, where MMR signifies “Match Making Rating”. The rest of the top 100 were mostly conspiracy subreddits — there were only 880 posts in parenting or medical-related subreddits. A more specific query looked for subreddits with submissions containing either “MMR shot”, “MMR jab”, “MMR injection”, “MMR vacc%”, or “MMR vax%”. Only 582 subreddits were returned, for a total count of 3,415 submissions. Of these, only 349 submissions belonged in subreddits that were devoted to parenting or medical questions, and even here the top-rating subreddits were conspiracy-related.

When the labelled 41,600 records of downloaded vaccine-related submissions were examined for VAEM-related discussions of MMR, only 5 posts were found where an MMR vaccine was attributed to an VAEM. Most of the discussions were around controversies relating the MMR vaccine to autism and bowel disease. An example of a post *mentioning* MMR is quoted below in Table 57, as it is a good example of the nature of the discussions found on Reddit, which tend to be lengthy and diverse. Although revealing in their detail, because of both their length and infrequency, posts like this are not suitable for the measurement of trends.

The vaccine that a reaction is attributed to seems to have been one of DTP, Hib or PCV13 (*“no more TDaP, no more HiB, no more PCV13”*), but extracting that information while verifying the lack of an explicit MMR link to the reaction is not the direction of this research.

Table 57: Reddit submission example

I now get to be 'that parent' the one with the under vaccinated child.

Today marked my son's 6 month well child - even though he was 6 and a half months old. Today he got only one vaccine and after speaking with the doctor this will be his last one until he's at least 5. Possibly older... possibly never have another vaccine for the rest of his life.

At 2 months I was suffering from severe anxiety. Although at the time I would have told you it was hormones I now accept that it was because I almost died, and my son almost died in the birthing process. The statistics for what we went through are very humbling. 5% of babies survive a full abruption. That's it. He is one of the 5%. We're very lucky and a fast and efficient medical team is who I have to thank for his very being.

So I had anxiety. I insisted on an alternate vaccine schedule to ease my worries that my own vaccine reactions as a baby would not pass on to him. Every one told me he would be fine, that those kinds of things don't happen. I believed them but insisted anyway. The altered schedule I'd drawn up only delayed one important vaccine by 1 month. It wasn't that big of a deal. Yes we were skipping the Hep A and Hep B altogether but other than that, we were going full bore with the rest.

He had a reaction.

He had a high fever and was screaming. Not fussy, not a bit 'off'... he screamed in the same way as he had when they had jabbed him and he wasn't stopping. Some Tylenol, some cuddling, eventually he fell asleep.

At 3 months he got sick. Bronchitis. No vaccinations.

At 4 months we repeated the ones he'd gotten at 2 months. This time, this time it was worse. A lot worse. There was no consoling him, he wouldn't eat, he wouldn't be distracted, he wouldn't sleep. Hour after hour after hour he screamed, and cried, and screamed. His fever wouldn't go down. I cried. My mother made me leave the house for 20 minutes convinced it was me that was causing his distress. He screamed louder when I wasn't there and my heart broke. He vomited violently, he got a rash. Against my instinct, against everything, my mother refused to take him to the ER, said I was overreacting, he'd calm down eventually. I had no car. I was trapped.

His fever lasted for 2 days before breaking. He was lethargic but improving slowly. In a stroller out for a walk he was content. That was the only place, so we took five walks that day.

At 5 months my doctor said - 1 shot, one we hadn't given him yet. Not the three we had planned at 3 months. Just 1, just the most important 1. He hadn't gotten one yet of that kind and maybe he'd be ok.

I will say this, it was better than the others. It wasn't good but it was doable. No screaming. No vomiting. Just lethargic and fussy but that could have been any day. I was just sensitive to it because of what happened the month before. I admit that. I admitted it at the time.

Today I got to see a specialist in communicable diseases. He looked at the pictures of the rash, he spoke to me at length about baby boy's reaction. We spoke about my anxiety and how I was feeling much better now so it was no longer an issue and he said something to me I didn't expect. He said to me, 'you don't have to worry anymore, your son won't be getting any more vaccinations until he's at least 5.'

I hadn't expected that. I had expected a lecture. I had expected to be told that we'd repeat the one that didn't really have a reaction. That I would need to just do one at a time until we were done to 'rule out the bad ones'. Nope. He told me about measles outbreaks, he told me about what to look for and how to keep my son safe as an un-vaccinated baby. How to keep other babies safe FROM him. He talked about titres and immunity all the while my pediatrician sat there and smiled. He told me about quarantine. I will say this, I feel amazingly relieved.

I am not anti-vaccination. I had expected to go through the entire series in some altered form to avoid reactions, not to be told that we were done. Now, nothing. No MMR, no Hep A, no Hep B, no polio, no flu, no more TDaP, no more HiB, no more PCV13.

I think I'll explain this to people by not telling them ever at all. All my friends and family will say I'm overreacting but I'm doing this under the guidance and recommendation of my pediatrician who is VERY pro-vaccination. Thanks for reading, I just really wanted to type that out but I'm not a blogger so I have no blog to add it to.

TL;DR - I'm pro-vaccination but due to bad reaction to them we're now a member of the under vaccinated crowd. The specialist says he can't recommend starting up again until he's at least 5 and then only one at a time.

Appendix J
Classification model definitions

The table below presents parameters and architectures of the classification models used in this study. For the traditional models, the vectorization method and parameters are also presented. For the traditional models, only the specific parameters the author used are presented. For the neural networks trained from scratch there are a few standard settings, such as optimizer and learning rate, which were used throughout. The Transformers were used with their defaults throughout, the key values from their configurations are presented. The ULMFiT model was also extensively tested, to try and improve the underlying language model's capacity to predict VAEM texts, but these experiments consisted of unlocking layers and fine-tuning them while evaluating the optimum number of iterations, rather than adjusting many parameters. For all the neural networks, a lot of the experimentation was to assess the optimum number of epochs and minibatches to get the best from the model, which was usually just before the models started overfitting — and was assessed based on changes in validation loss, and, after training, on test F1-Scores. For the models that were trained from scratch, many experiments were conducted to arrive at these settings, but the detail of these is not presented.

Table 58: Model definitions and parameters

Model	Model Definition	Vectorizer Definition
Logistic Regression CV	LogisticRegressionCV(Cs=50, max_iter=2000, random_state=23)	TfidfVectorizer(sublinear_tf=True, max_df=0.5, ngram_range = (1, 2), use_idf=False)
Stochastic Gradient Descent Classifier	SGDClassifier(alpha=0.0001, max_iter=50000, penalty='l2', tol=0.001, random_state=23)	TfidfVectorizer(sublinear_tf=True, max_df=0.5, token_pattern ='(?ui)\\b\\w*[A-Za-z]{2,}\\w*\\b', ngram_range=(1, 2), use_idf=True)
Linear Support Vector Machines	LinearSVC(C=1, tol=0.001, random_state=23)	TfidfVectorizer(sublinear_tf=True, binary=False, ngram_range=(1, 2), use_idf=True)

Random Forest Classifier	RandomForestClassifier(n_estimators=1000, max_features=10, max_depth=None, min_samples_leaf=1, min_samples_split=3, criterion='entropy', bootstrap=False, oob_score=False, random_state=23)	TfidfVectorizer(max_df=0.5, sublinear_tf=True, token_pattern=u'(?ui)\\b\\w*[A-Za-z]{2,}\\w*\\b', use_idf=False)
Extra Trees Classifier	ExtraTreesClassifier(n_estimators=1000, max_features=10, max_depth=None, min_samples_leaf=1, min_samples_split=3, criterion='entropy', bootstrap=False, oob_score=False, random_state=23)	TfidfVectorizer(max_df=0.5, sublinear_tf=True, token_pattern=u'(?ui)\\b\\w*[A-Za-z]{2,}\\w*\\b', use_idf=False)
Multinomial Naïve Bayes	MultinomialNB(alpha=0.15, class_prior=None, fit_prior=False)	TfidfVectorizer(sublinear_tf=True, max_features = None, max_df=0.5, ngram_range = (1, 2), token_pattern = '(?ui)\\b\\w*[A-Za-z]{2,}\\w*\\b', use_idf=False)
Naïve Bayes SVM	NBSVM(C=1, alpha = 0.01, beta=1)	TfidfVectorizer(sublinear_tf=True, binary = False, norm = 'l1', max_features = 10000, max_df=0.5, ngram_range = (1, 3), use_idf=False)

XGBoost	XGBClassifier(learning_rate=0.04, n_estimators=900, colsample_bytree=0.6, gamma=1, max_depth=5, min_child_weight=1, subsample=0.6, objective='binary:logistic', random_state=23)	TfidfVectorizer(max_df=0.5, sublinear_tf=True, use_idf=True)
All Neural Networks trained from scratch	activation : selu optimizer : AdamW learning_rate : 0.001 init_weight : True init_weight_value : 2.0 optim_momentum_value : 0.9 batch_normalizations : False clip : 5 weight_decay : 1e-8 batch_size : 32	
CNN	[Conv2d(1, 100, kernel_size=(1, 100), stride=(1, 1), bias=False), Conv2d(1, 100, kernel_size=(2, 100), stride=(1, 1), padding=(1, 0), bias=False), Conv2d(1, 100, kernel_size=(3, 100), stride=(1, 1), padding=(1, 0), bias=False)] CNN_Text((embed): Embedding(4882, 100, padding_idx=1, scale_grad_by_freq=True) (dropout): Dropout(p=0.5, inplace=False) (dropout_embed): Dropout(p=0.1, inplace=False) (fc): Linear(in_features=300, out_features=2, bias=True))	
CNN-BiLSTM	[Conv2d(1, 100, kernel_size=(1, 100), stride=(1, 1)), Conv2d(1, 100, kernel_size=(2, 100), stride=(1, 1), padding=(1, 0)), Conv2d(1, 100, kernel_size=(3, 100), stride=(1, 1), padding=(1, 0))] CNN_BiLSTM((embed): Embedding(4882, 100, padding_idx=1) (bilstm): LSTM(100, 300, num_layers=2, dropout=0.5, bidirectional=True) (hidden2label1): Linear(in_features=900, out_features=450, bias=True) (hidden2label2): Linear(in_features=450, out_features=2, bias=True) (dropout): Dropout(p=0.5, inplace=False))	

CNN-BiGRU	[Conv2d(1, 100, kernel_size=(1, 100), stride=(1, 1)), Conv2d(1, 100, kernel_size=(2, 100), stride=(1, 1), padding=(1, 0)), Conv2d(1, 100, kernel_size=(3, 100), stride=(1, 1), padding=(1, 0))] CNN_BiGRU((embed): Embedding(4882, 100, padding_idx=1) (bigru): GRU(100, 300, num_layers=2, dropout=0.5, bidirectional=True) (hidden2label1): Linear(in_features=900, out_features=450, bias=True) (hidden2label2): Linear(in_features=450, out_features=2, bias=True) (dropout): Dropout(p=0.5, inplace=False))
CNN-LSTM	CNN_LSTM((embed): Embedding(4882, 100, padding_idx=1) (dropout): Dropout(p=0.5, inplace=False) (lstm): LSTM(100, 300, num_layers=2, dropout=0.5) (hidden2label1): Linear(in_features=600, out_features=300, bias=True) (hidden2label2): Linear(in_features=300, out_features=2, bias=True))
LSTM	LSTM((embed): Embedding(4882, 100, padding_idx=1) (lstm): LSTM(100, 300, num_layers=2, dropout=0.5) (hidden2label): Linear(in_features=300, out_features=2, bias=True) (dropout): Dropout(p=0.5, inplace=False) (dropout_embed): Dropout(p=0.1, inplace=False))
BiLSTM	LSTM(100, 150, bias=False, dropout=0.5, bidirectional=True) BiLSTM((embed): Embedding(4882, 100, padding_idx=1) (bilstm): LSTM(100, 150, bias=False, dropout=0.5, bidirectional=True) (hidden2label1): Linear(in_features=300, out_features=150, bias=True) (hidden2label2): Linear(in_features=150, out_features=2, bias=True))
GRU	GRU((embed): Embedding(4882, 100, padding_idx=1) (gru): GRU(100, 300, num_layers=2, dropout=0.5) (hidden2label): Linear(in_features=300, out_features=2, bias=True) (dropout): Dropout(p=0.5, inplace=False))
BiGRU	BiGRU((embed): Embedding(4882, 100, padding_idx=1) (bigru): GRU(100, 300, num_layers=2, dropout=0.5, bidirectional=True) (hidden2label): Linear(in_features=600, out_features=2, bias=True) (dropout): Dropout(p=0.5, inplace=False))

All Transformers	max_seq_length : 64 learning_rate : 2e-5 batch_size : 32, or 16 for larger models iterations per epoch : length training data / batch size adam_epsilon : 1e-8 warmup_steps : 0 max_grad_norm : 1.0 random_seed : 42

BERT	{
	"attention_probs_dropout_prob": 0.1,
	"hidden_act": "gelu",
	"hidden_dropout_prob": 0.1,
	"hidden_size": 768,
	"initializer_range": 0.02,
	"intermediate_size": 3072,
	"layer_norm_eps": 1e-12,
	"max_position_embeddings": 512,
	"num_attention_heads": 12,
	"num_hidden_layers": 12,
	"output_attentions": false,
	"output_hidden_states": false,
	"output_past": true,
	"type_vocab_size": 2,
	"vocab_size": 30522
	}

RoBERTa	{
	"attention_probs_dropout_prob": 0.1,
	"hidden_act": "gelu",
	"hidden_dropout_prob": 0.1,
	"hidden_size": 768,
	"initializer_range": 0.02,
	"intermediate_size": 3072,
	"layer_norm_eps": 1e-05,
	"max_position_embeddings": 514,
	"num_attention_heads": 12,
	"num_hidden_layers": 12,
	"output_attentions": false,
	"output_hidden_states": false,
	"output_past": true,
	"type_vocab_size": 1,
	"vocab_size": 50265
	}

RoBERTa Large	{
	"attention_probs_dropout_prob": 0.1,
	"hidden_act": "gelu",
	"hidden_dropout_prob": 0.1,
	"hidden_size": 1024,
	"initializer_range": 0.02,
	"intermediate_size": 4096,
	"layer_norm_eps": 1e-05,
	"max_position_embeddings": 514,
	"num_attention_heads": 16,
	"num_hidden_layers": 24,
	"output_attentions": false,
	"output_hidden_states": false,
	"output_past": true,
	"pruned_heads": {},
	"torchscript": false,
	"type_vocab_size": 1,
	"vocab_size": 50265
	}

XLNet	{
	"attn_type": "bi",
	"clamp_len": -1,
	"d_head": 64,
	"d_inner": 3072,
	"d_model": 768,
	"dropout": 0.1,
	"end_n_top": 5,
	"ff_activation": "gelu",
	"initializer_range": 0.02,
	"layer_norm_eps": 1e-12,
	"n_head": 12,
	"n_layer": 12,
	"n_token": 32000,
	"output_attentions": false,
	"output_hidden_states": false,
	"output_past": true,
	"start_n_top": 5,
	"summary_activation": "tanh",
	"summary_last_dropout": 0.1,
	"summary_type": "last",
	"summary_use_proj": true,
	"untie_r": true,
	}

XLNet Large	{
	"clamp_len": -1,
	"d_head": 64,
	"d_inner": 4096,
	"d_model": 1024,
	"dropout": 0.1,
	"end_n_top": 5,
	"ff_activation": "gelu",
	"initializer_range": 0.02,
	"layer_norm_eps": 1e-12,
	"n_head": 16,
	"n_layer": 24,
	"output_attentions": false,
	"output_hidden_states": false,
	"output_past": true,
	"start_n_top": 5,
	"summary_activation": "tanh",
	"summary_last_dropout": 0.1,
	"summary_type": "last",
	"summary_use_proj": true,
	"vocab_size": 32000
	}

XLM	{
	"attention_dropout": 0.1,
	"dropout": 0.1,
	"emb_dim": 2048,
	"init_std": 0.02,
	"layer_norm_eps": 1e-12,
	"mask_index": 5,
	"max_position_embeddings": 512,
	"n_heads": 16,
	"n_layers": 12,
	"output_attentions": false,
	"output_hidden_states": false,
	"output_past": true,
	"pad_index": 2,
	"start_n_top": 5,
	"use_lang_emb": true,
	"summary_first_dropout": 0.1,
	"summary_proj_to_labels": true,
	"summary_type": "first",
	"summary_use_proj": true,
	"vocab_size": 30145
	}

ULMFiT	learning rate = 1e-3 optimizer = Adam(), betas=(0.9, 0.99) loss function = FlattenedLoss of CrossEntropyLoss() callbacks = [RNNTrainer learn: ... alpha: 2.0 beta: 1.0]