Social Media Analysis on COVID-19 Vaccine

COMP61332 Text Mining Coursework 2
Xinyi Ouyang, Chuhan Qiu, Zhangli Wang, Mingchen Wan, Mochuan Zhan

Abstract

In recent years, the use of social media has been tremendous increase and Social Media Analytics (SMA) has become a significant area in information systems research (Stieglitz et al., 2014). In this project, we applied SMA pipeline to COVID-19 vaccine to capture the public's response to the vaccination. More specifically, we analysed public's response from different perspectives through sentiment analysis, topic modeling and named entity recognition.

1 Background & Introduction

Infectious diseases cause a high proportion of deaths (Andre et al., 2008) and have a tremendous influence on human society (Zhang et al., 2017). Social media, such as Twitter, has become pervasive and provides a public communicating platform during the pandemic. According to Qazi et al. (2020), it is necessary for researchers and the government to understand public's opinion toward the epidemic and capture the topics from social media to grasp the situational information.

In the opinion of Andre et al. (2008), vaccination can be regarded as an effective tool for public health during the pandemic and public's opinion and willingness towards vaccination are extremely important (Yousefinaghani et al., 2021).

Since December of 2019, COVID-19 has swept across the world and has had serious impacts on all aspects of human society (Harper et al., 2020). This project aimed at finding out the public response towards the COVID-19 vaccines in January 2021 by implementing a Social Media Analytics pipeline on Twitter. More specifically, the main objectives were : (1) to identify the public sentiment on the vaccination, (2) to find the vaccines' manufacturers they most like to talk about, and (3) to capture the potential topics when they talk about the COVID-19 vaccines.

2 Related Work

There has been a massive increase in social media usage since the COVID-19 outbreak(Qazi et al., 2020), and social media data has been utilized to capture people's opinions on COVID-19 vaccination in several studies (DeVerna et al., 2021). The research of Piedrahita-Valdés et al. (2021) used a hybrid model to analyze the sentiment polarity of the COVID-19 vaccine, DeVerna et al. (2021) evaluated the public's attitude towards vaccination, and Nuzhath et al. (2020) identified vaccination topics discussed in Twitter by using topic modeling with LDA. Therefore, Twitter data has been proved to be an important tool for infodemiology studies (Chew and Eysenbach, 2010).

The study of Yousefinaghani et al. (2021) tracked the hashtags, keywords, and main themes in tweets with sentiment polarity, to analyze the evolution of sentiments, capture the keywords and themes, and compare different vaccine manufacturers. Sentiment analysis, keywords extraction, and topic modeling were used to gain the public's opinions, and besides the text data, other information such as location and users were also used for analysis.

The above study is highly similar to our research questions and shows the potential for analysis of Twitter data to help public health agencies take positive responses. In our project, we referred to the above project, applying sentiment analysis, Named Entity Recognition, and topic model on the Tweets related to COVID-19 vaccination, to figure out the public opinion towards vaccines and achieve the objectives.

3 Methodology

3.1 Data collection

Twitter provides an official API for developers to access Twitter data, and Tweepy library can help developers to collect Tweets by entering keywords, usernames, locations or the combination. But with

the epidemic out of sight, people talk less about COVID-19 nowadays. In this case, the data in 2021 is more convincing if we want to research on public responses to COVID-19 vaccines. So we used an open-source dataset in Kaggle instead of collecting Tweets by ourselves. It contains around 380,000 records from January 8th, 2020 with the keyword 'CovidVaccine'. In this project, we selected 48510 Tweets from January 1st to January 28th in 2021 as our analysis data.

3.2 Preprocessing

The preprocessing techniques applied for this project are as follows:

1. Remove noisy and undesired instances.
2. Filter tweets posted in January 2021.
3. Replace contractions.
4. Replace negations with their antonyms.
5. Lemmatize all words.
6. Demojize emojis.
7. Remove stop words.
8. Remove URLs, newlines, punctuations, and usernames.
9. Lowercase all characters and remove duplicated instances.

The general purpose of natural language preprocessing is to transform the original texts to the forms that SMA models can extract the meaning of words or sentences more comprehensively. Specifically, preprocessing step 3 splits word such as "isn't" to "is not" for retrieving correct tokens in the following steps. The followed-up step 4 captures "not" tokens in sentences and their succeeding words and transforms them with the antonyms of the succeeding words. This process aims to involve and embed the negative meanings to the output. Step 5 aims to restitute words with the same meanings to the same forms to lower the difficulty of training SMA models. Step 6 is to transform emojis to interpretable forms for the models. Step 7 is to remove stop words that are considered less contributable to meaningful semantics for SMA.

3.3 NLP Techniques

3.3.1 Sentiment Analysis

In order to explore people's attitudes and sentiment towards COVID-19 vaccine, such as positive, negative or neutral, this project used Vader sentiment as an analysis tool to analyze every text in the data set.

SentimentIntensityAnalyzer output for (pos, neu, neg, compound) discriminant criteria for the project will compound >= 0.05 as positive, compound > -0.05 and compound < 0.05 as neutral, Compound <= -0.05 was considered negative.

For the results, a holistic analysis was first performed to obtain the proportion of different emotions about COVID-19 vaccine in the entire text data set. To gauge people's attitudes and emotions towards the introduction and development of COVID-19 vaccines. Then, the January data were analyzed day by day to explore whether people's mood fluctuated significantly from day to day. If so, the causes could be found, such as whether important people were infected with COVID-19 and policy changes.

3.3.2 Topic Modelling

Topic modelling is an approach to automatically finding a series of topics from a collection of documents. In this project, we need to draw several topics from Twitter's data to realize the main components of what people think about COVID-19 vaccines. The main process is divided into the following steps: (1) Clean the data (i.e. removing stopwords, making bigrams, and lemmatizing). (2) Create corpus and dictionary by the use of BoW representation. (3)Train the Latent Dirichlet Allocation (LDA) model.

LDA could assign topics to documents and generate topic distributions based on a given collection of text (Petterson et al., 2010). Key steps of LDA to approximate these distributions: (1) Select K, the number of topics present. (2) Go through each document, and randomly assign each word to one of K topics. (3) Iterate through each document. For each document, go through each word and reassign a new topic, where we choose topic t with a probability p(topic t | document d) * p(word w | topic t) based on the last round's distribution. (4) Keep iterating until topic assignments reach converge.

3.3.3 Named Entity Recognition

To find out the vaccine brands that the public is concerned about most, we utilized NER (Named Entity Recognition) to identify vaccine-related terms. We analyze the data annotated with named entities in two ways: 1. Count the number of times all vaccines appeared in January and rank them based on their occurrences. 2. Separate data into weeks and plot the distribution of occurrences of different covid-vaccines.

Because existing models off-the-shelf are not

vaccine-targeted and only a minority of vaccine brands are included, therefore we trained our model based on the structure of Spacy (Honnibal and Montani, 2017).

In order to train the vaccine-targeted model, we annotated a number of tweets mentioning vaccine brands to build up a dataset for training and testing, the potential brands including Pfizer, Moderna, AstraZeneca, Covaxin, Sputnik V, Sinopharm ,and Sinovac. The accuracy of our model turns out to be good, by manually checking, most vaccine brands in tweets are correctly annotated.

4 Visualization & Analysis

4.1 Sentiment Analysis to Vaccine

For the results of sentiment analysis, the project adopts the analysis of overall data and day-by-day analysis.

As shown in Figure 1. The results of the analysis basically met expectations, and multiple COVID-19 vaccines were released and launched in January 2021. After the panic and anxiety of the novel coronavirus infection. About 47 percent of the population reported positive feelings about the introduction of the vaccine, and nearly half reported positive feelings when referring to the COVID-19 vaccine. About 30 percent of the population is neutral on vaccines, which is probably a group of people who do not have a deep interest in novel Coronavirus and vaccines, believing that novel Coronavirus is just an ordinary infectious disease. A minority of about 20 percent have negative feelings about COVID-19 vaccines, most of them from anti-vaccine groups, according to the comments.

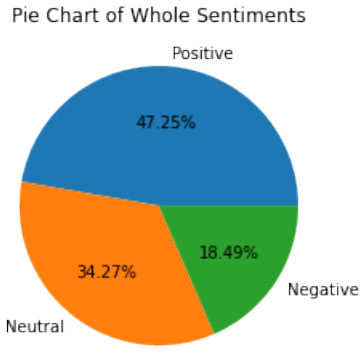

Figure 1: Proportion of Different Sentiments

As shown in Figure 2, it can be seen that the proportion of the three sentiment is basically the same every day. Positive sentiment are always dominant and negative sentiment are less. Can be seen from the diagram in mid-January and one end of the month, the crowd for new vaccine the discussion of this topic volume increased significantly, as a result, the team take look for the two periods have related events, after a search, found in mid-January new confirmed cases of obvious increase rapidly, possible reason is that the delta mutant strains spread quickly.

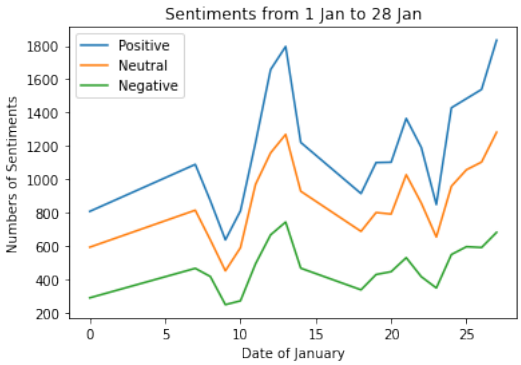

Figure 2: Distribution of Sentiments in January

4.2 Topics Analysis

In order to find better result for Topic Modelling, we first used CV Coherence Score to help us find the suitable number of topics. As shown in the Figure 3 below, we calculated coherence scores on the scale of 2 to 15 topics. Since the coherence score seems to increase as the number of topics increases, it may make more sense to choose a model that gives the highest CV before flattening out or dropping significantly. Therefore, 13 should be chosen as the number of topics to train our model.

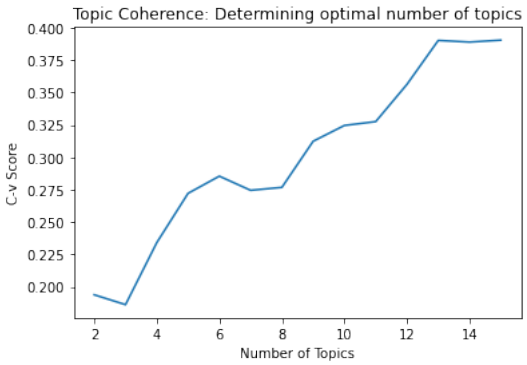

Figure 3: Topic Coherence Scores in Different Number of Topics

The result of Topic Modelling can be found in Table 1, which shows 13 topics people were talking

Topic	Relevancy λ and terms
1	λ=1;healthcare,way,need,next,rollout,phase,end,line,county,whole
2	λ=1;great,uk,well,news,community,really,yet,elderly,anyone,use
3	λ=1;thank,plan,distribution,virus,astrazeneca,effective,video,forward,access,new
4	λ=1;syringe,staff,nhs,receive,proud,go,clinic,never,first,give
5	λ=1;health,new,public,right,eligible,enough,someone,call,canada,risk
6	λ=1;important,science,information,appoinment,free,number,big,checkmarkbutton,far,mom
7	λ=1;first,dose,second,pfizer,yesterday,shot,already,old,morning,nurse
8	λ=1;good,grateful,coronavirus,arm,shot,hospital,keep,new,news,team
9	λ=0.2;last,wait,full,hear,finally,long,stop,lockdown,research
10	λ=1;india,world,january,supply,covishield,pm,read,covaxin,group,drive
11	λ=1;safe,part,much,government,hope,state,family,home,join,frontline
12	λ=1;care,start,pandemic,update,health,country,little,biden,watch,doctor
13	λ=0.5;many,let,happy,mederna,everyone,back,soon,know,vaccinate,question

Table 1: Topics about Covid Vaccine in Jan 2021

about on Twitter in January 2021. For each topic, we have listed the 10 terms that are most relevant to each topic with different settings of λ. And these terms are listed in descending order of their weight in topics. For the most part, these topics are expected, except for those that arise due to ongoing events in the real world. For instance, there was a vaccine death in India in January 2021, which led to one of our topics involving India and 'Covishield' which is the name of a serum institute in India.

What we glean from these topics is expectation and uncertainty about what people think about vaccines. This could be seen as by the first topic, the most popular one, vaccine-related terms like 'rollout' and 'phase' got more mention. However, in the topics below, people have shown a variety of attitudes towards vaccines. For example, the second topic is about vaccines having 'great' and 'well' effects on the 'elderly', but people also commented on the long wait for vaccines in the eighth topic.

4.3 Vaccine Named Entity Analysis

Although our NER model has high accuracy in recognizing COVID-19 vaccine brands, the side effect is obvious. It classifies misspelled words into vaccine brands so that the result contains a group of low occurrence words. To avoid accepting these misspelled words, only the vaccine brands with the highest occurrences were kept.

Fig.4 demonstrates the top 12 covid vaccine brands that have the highest occurrence in January. Among all covid vaccine brands shown in this figure, Pfizer takes the most public attention on Twitter, which is 33.93%, while Sinovac only obtains 1.41% of vaccine topics. The other three popular vaccine brands Moderna, AstraZeneca, and Convaxin occupy 21.07%, 11.92%, and 8.66% respectively.

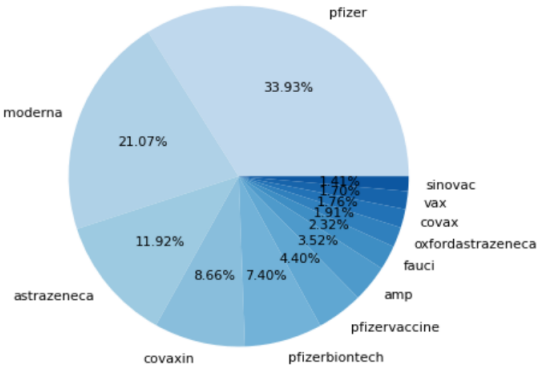

Figure 4: Top 12 Covid Vaccine Brands in Jan 2021

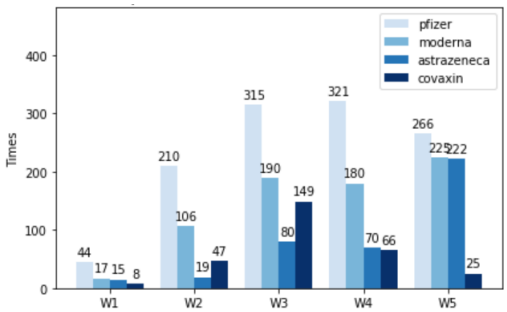

Figure 5: The Distribution of 4 Popular Covid Vaccines' Occurrence In Jan

In further analysis, we plotted the distribution of 4 Popular Covid vaccines' occurrences in Fig.5. The data is not evenly distributed, with significantly

fewer data in the first two weeks. Two possible explanations are as follows. Firstly, the data is split into five weeks according to the calendar and the first week only has three days (1st Jan, 2nd Jan, 3rd Jan). In addition, the data during 2nd Jan and 7nd Jan were missing. In this case, week 1 only has data for a single day, and week 2 has data for less than half a week. Ideally, if we scale the data for the first two weeks, the data is roughly evenly distributed.

However, there exists two exceptions. The occurrence of Covaxin and AstraZeneca reached unusual levels on week 3 and week 5 respectively. With news regrading evidence,the possible reason for Covaxin's high occurrence might be the news about the death of the Covaxin volunteer on 9th Jan 2021. While the possible reason for AstraZeneca might be the topics about EU threatening to block Covid vaccine exports amid AstraZeneca shortfall on 25th Jan 2021.

References

Francis E Andre, Robert Booy, Hans L Bock, John Clemens, Sibnarayan K Datta, Thekkekara J John, Bee W Lee, S Lolekha, Heikki Peltola, TA Ruff, et al. 2008. Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bulletin of the World health organization*, 86:140–146.

Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

Matthew DeVerna, Francesco Pierri, Bao Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Cristopher Torres-Lugo, Kai-Cheng Yang, Fil Menczer, and John Bryden. 2021. Covaxxy: A global collection of english twitter posts about covid-19 vaccines. *arXiv e-prints*, pages arXiv–2101.

L Harper, N Kalfa, GMA Beckers, M Kaefer, AJ Nieuwhof-Leppink, Magdalena Fossum, KW Herbst, D Bagli, ESPU Research Committee, et al. 2020. The impact of covid-19 on research. *Journal of pediatric urology*, 16(5):715–716.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Tasmiah Nuzhath, Samia Tasnim, Rahul Kumar Sanjwal, Nusrat Fahmida Trisha, Mariya Rahman, SM Farabi Mahmud, Arif Arman, Susmita Chakraborty, and Md Mahbub Hossain. 2020. Covid-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of twitter data.

James Petterson, Wray Buntine, Shravan Narayanamurthy, Tibério Caetano, and Alex Smola. 2010. Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Hilary Piedrahita-Valdés, Diego Piedrahita-Castillo, Javier Bermejo-Higuera, Patricia Guillem-Saiz, Juan Ramón Bermejo-Higuera, Javier Guillem-Saiz, Juan Antonio Sicilia-Montalvo, and Francisco Machío-Regidor. 2021. Vaccine hesitancy on social media: Sentiment analysis from june 2011 to april 2019. *Vaccines*, 9(1):28.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15.

Stefan Stieglitz, Linh Dang-Xuan, Axel Bruns, and Christoph Neuberger. 2014. Social media analytics. *Business & Information Systems Engineering*, 6(2):89–96.

Samira Yousefinaghani, Rozita Dara, Samira Mubareka, Andrew Papadopoulos, and Shayan Sharif. 2021. An

analysis of covid-19 vaccine sentiments and opinions on twitter. *International Journal of Infectious Diseases*, 108:256–262.

Xinhong Zhang, Daqing Jiang, Tasawar Hayat, and Bashir Ahmad. 2017. Dynamics of a stochastic sis model with double epidemic diseases driven by lévy jumps. *Physica A: Statistical Mechanics and its Applications*, 471:767–777.