Analysing Social Media Discourse on COVID-19 Vaccine

Aparna(11106479) and Balendra Singh(11166812) and Mariam Jamilah(11106105)
and Surya Ramessh(11041102) and Thushara Nair(11078927)
University of Manchester

Abstract

This research paper utilizes social media analysis techniques, such as sentiment analysis, named entity recognition, and topic modeling, to investigate how people perceive the COVID-19 vaccine using data from Twitter. The primary aim of the study is to examine the general sentiment towards COVID-19 vaccination and the most commonly discussed topics related to it. Furthermore, the study aims to explore whether there have been any changes in sentiment over time. The research shows that initially, people were positive towards the vaccine, as it was seen as a promising solution for the future. However, sentiment later shifted to neutrality, possibly due to uncertainty surrounding the vaccine's effectiveness. Negative emotions were also observed, particularly among individuals who experienced flu-like symptoms after getting vaccinated. Social media discussions around this topic were mostly negative. The study's findings provide valuable insights into public perceptions of the COVID-19 vaccine and can aid in the development of effective vaccination campaigns and communication strategies.

1 Introduction

Social media analytics (SMA) entails the creation and evaluation of informatics tools for collecting, monitoring, analysing, summarising, and visualising social media data in order to extract useful trends and intelligence.(1) SMA provides valuable feedback regarding the perspectives of individuals, which can influence future development decisions.(2)

The COVID-19 pandemic has sparked a range of emotions and opinions, from fear and anxiety to hope and optimism. While some people worldwide were optimistic about the return to normalcy after vaccine introduction, others are uncertain about the safety and efficacy due to a diversity of beliefs, political perspectives, and personal experiences.(3) SMA can help us understand how these emotions and opinions are reflected in social media conversations.

Twitter being one of the most popular social media platforms, with millions of active users posting and sharing information in real-time. Twitter data can provide a wealth of information about public opinion on the COVID-19 vaccine, making it an essential dataset for SMA.(4)

This research paper will analyse the Twitter data on the COVID-19 vaccine, to gain a better understanding of people's opinions and concerns about the COVID-19 vaccine and how they have evolved over time. This information can be useful to policymakers, public health officials, and healthcare providers in shaping their communication strategies and addressing people's concerns about the vaccine.

2 Related Work

Social media analytics is crucial to understand the topic of discussion on a subject and how it is being perceived, thereby identifying the influencing factors and key demographics.(5) Due to the widespread use of social media and the low entry barrier for posting a message, sentiment mining focuses on different perspectives such as task-oriented, granularity-oriented, methodology-oriented.(6) A new method that coupled context-based topic modelling on top of opinion mining enabled better recognition of emerging social attitudes and the interpretation of public responses on the Australian Federal Election 2010.(7) Another research proposed a neural-based NER approach because texts in social networks remain challenging, with entities representing a small proportion of proper names, making it difficult to generalise(8), and social media texts not adhering to syntax rules.(9)

Research Questions

1. *What is the dominant attitude towards COVID-19 immunization? Have there been any shifts in sentiment over time?*

2. *What are the most commonly addressed subjects regarding COVID-19 vaccination, and how have they developed over time?*

3. *Which persons, nationalities, religious/political groups, organisations and geopolitical entities were discussed in the dominant topics?*

3 Methodology

In order to investigate the research questions pertaining to social media analytics on COVID-19 vaccines, we developed a comprehensive experimental framework that consisted of four distinct phases. The first phase involved collecting and preprocessing data. Subsequently, we conducted sentiment analysis on the preprocessed data using various techniques to identify the most effective method of categorizing the data as positive, negative, or neutral. The labeled data was then utilised to address the primary research question and explore subsequent questions. The third phase entailed utilising the labeled data to perform topic modeling via , which enabled us to identify the top 10 dominant topics discussed within each of the distinct class labels. Lastly, we employed spaCy to perform named entity recognition, which allowed us to extract the relevant entities that pertained to the objectives of the study.

3.1 Dataset Description and Preprocessing

Only 1% of the data from Twitter is accessible to us through the APIs due to rate limiting.(10) Therefore, the "snscrape" Python library was employed to scrape tweets through Twitter's API without restrictions or request limits. Tweets were scraped using the keywords between the timelines. To determine if there has been a change in sentiment from the past, we utilise historical data from Github and Kaggle that will allow for a comprehensive analysis.

A total of eight sources were combined, after which redundant records and less-important features were removed. The final dataset consisted of around 150 thousand records with 8 essential columns, including the tweet's description, date, retweet, user name, etc.

The present study undertakes a series of data preprocessing steps to clean and transform Twitter data related to the COVID-19 vaccine for application in sentiment analysis, topic modeling, and named entity recognition. These steps involve the removal of URLs, Twitter-specific syntax, common abbreviations, non-alphabetic characters, short tweets, and stop words. Additionally, tokenization, lemmatization, and detokenization are implemented to reduce the sparsity and noise in the text data. The removal of duplicates, on the assumption that they are from the same users, is also carried out. The overarching objective of these preprocessing measures is to enhance the accuracy of the analyses and render the data suitable for further employment in social media analytics related to the COVID-19 vaccine.

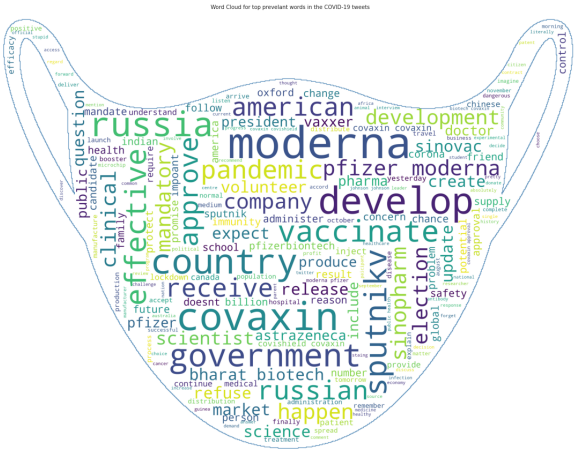

Figure 1: WordCloud of COVID-19 Vaccine Twitter Dataset

The visualization in Figure 1 displays a word cloud derived from the twitter dataset utilised in this study. It effectively conveys the overall context of the dataset and highlights the most commonly occurring words. Thus, enabling easy interpretation and comprehension of the dataset's content.

3.2 Sentiment Analysis

In this study, we carried out a sentiment analysis experiment to analyse Twitter data regarding the COVID-19 vaccine, utilising two different sentiment analyser: TextBlob and VADER. The polarity of each tweet was computed using TextBlob, and sentiment labels were assigned based on the manual inspection of the polarity score, with scores equal to 0 being considered neutral, scores less than 0 being negative, and scores greater than 0 being positive.

Similarly, VADER was used to calculate the sentiment of each tweet and to assign sentiment labels based on the compound score. Scores less than or equal to -0.5 were considered negative, scores greater than or equal to 0.5 were considered positive, and anything else was considered neutral.

The results of each approach were saved and compared using manual inspection and visualisations to determine which method was better suited for inferring people's sentiments over time and for further analysis.

3.3 Topic Modeling

In this research paper, we utilised BERTopic(11), a method that leverages pre-trained sentence embeddings and clustering models to conduct topic modeling in an efficient manner.

To accomplish this, we employed encoding to generate embeddings for the input document, which were subsequently used by the HDBSCAN clustering model to generate topics. For the topic modeling process, we used BERTopic in conjunction with a pre-trained sentence embedding model for the English language. To extract the resulting topics and their corresponding probabilities from the input document and embeddings, we set the 'calculate probabilities' parameter to true. Additionally, we filtered out small topics by setting the 'minimum topic size' parameter to 100, and we determined the minimum and maximum range for n-grams used in vectorisation by setting the 'n-gram range' parameter to (1,2).

3.4 Named Entity Recognition

The current study utilised the Named Entity Recognition (NER) approach, a commonly used technique in Information Extraction (IE), to detect and categorise particular entities in a corpus of text. The spaCy Natural Language Processing (NLP) library was utilised to perform the NER analysis, with a focus on identifying entities falling under the categories of PERSON, NORP, ORG, and GPE, which were deemed relevant to the study's objectives.

4 Results

Figure 2 shows the results of our sentiment analysis experiment which reveals that while both TextBlob and VADER produced similar positive sentiment labels, they differed in their labeling of negative and neutral sentiments. We conducted manual inspection of the unlabeled data to validate the sentiment labels, and the results indicated that VADER had more accurate labeling than TextBlob. VADER is more effective than TextBlob for sentiment analysis on Twitter datasets due to its ability to handle unique features of social media texts such as emojis, slang, and abbreviations. (12)

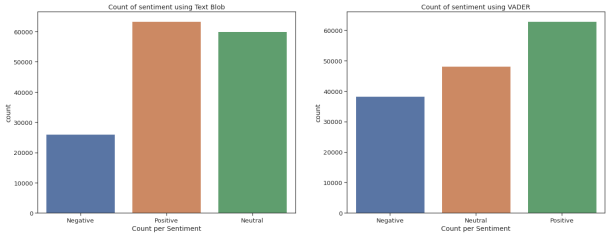

Figure 2: Comparison of Sentiment Analysis Approaches

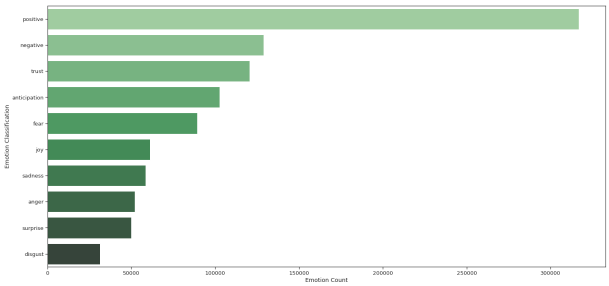

Figure 3: Dominant Sentiments towards COVID-19 Vaccine

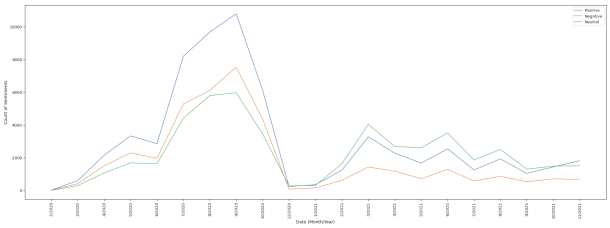

Figure 4: Sentiment Trends Over Time

The analysis of Figure 3 reveals a strong positive sentiment towards the development of vaccines, but upon closer examination, it becomes evident that there are also other prevalent emotions towards vaccines. In order to gain a deeper understanding, we conducted a detailed examination of the evolution of emotions over time and found that while vaccines were initially received positively, this sentiment underwent a change over time as depicted in Figure 4. It suggests that people initially responded positively to the COVID-19 vaccine upon its release, but negative emotions were also observed due to fear of the unknown and the possibility of

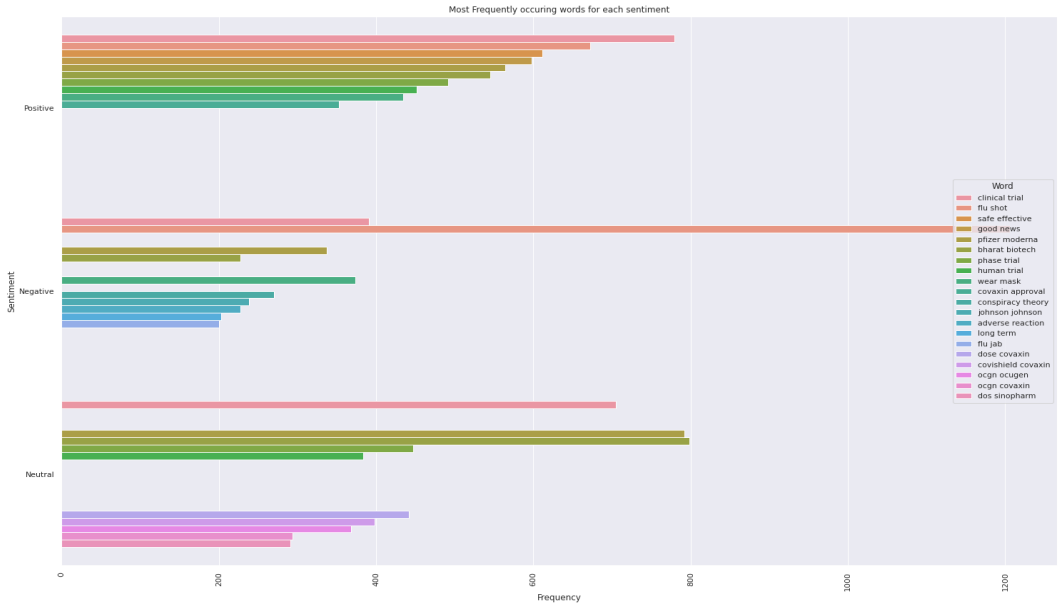

Figure 5: Most Frequently occurring words for each Sentiments

experiencing flu-like symptoms after vaccination. However, as time passed, people's opinions about the vaccine gradually shifted towards neutrality. The initial positive response may have been influenced by the anxiety and fear surrounding the pandemic, and the novelty of the vaccine. As more people received the vaccine, the excitement surrounding it decreased, leading to a more neutral outlook. This phenomenon is a common human behavior where initial novelty generates high interest, but over time, people tend to become accustomed to it, resulting in a decrease in interest.

Furthermore, according to the data presented in Figure 5, certain phrases were associated with specific emotions in the study. The term "clinical trial" was mostly associated with positive emotions, likely due to the sense of urgency related to developing a vaccine during the pandemic. Clinical trials gave people hope that the pandemic could be overcome. On the other hand, "flu shot" was primarily linked with negative emotions, indicating that some individuals may have developed unfavorable feelings after getting the flu despite being vaccinated. However, "flu shot" was also the second most frequently used phrase in a positive context, which could imply that people understand that experiencing the flu after vaccination might indicate that the vaccine is working effectively. In contrast, the phrases "Bharat Biotech," "Pfizer Moderna" were mainly used with a neutral emotional tone, indicating mixed feelings about the vaccines' efficacy and safety, which could be attributed to the initial uncertainty surrounding them.

The analysis of Figure 6 indicates that the main topic discussed across all three sentiments (positive, negative, and neutral) is related to vaccines, specifically Sinopharm and Covaxin. These findings support our previous conclusions that vaccines evoke a range of emotions, including fear and anxiety, as well as optimism and hope for the future. The presence of neutral sentiment could be due to uncertainty surrounding the effectiveness and safety of these vaccines. Furthermore, the intense emotions associated with the antivaxxer movement should also be noted. Those who oppose vaccines are often seen as prioritising their personal beliefs over the greater good of society, which generates disapproval from many individuals. Nevertheless, some people sympathize with antivaxxers' concerns about vaccine safety and individual rights, leading to a sense of support. These results suggest that the vaccine debate is a contentious and polarizing issue, eliciting strong and diverse emotional responses from different groups.

Additionally, Figure 7 illustrates the trend of dominant topics discussed in the research. The majority of these topics are related to various COVID-19 vaccines and the organizations that produced them. It is evident from the graph that the discussion about these vaccines reached its highest peak between January 2021, when they were initially introduced in the market, and July 2021, when a significant portion of the population had already received the vaccine, and the pandemic

Figure 6: Dominant Topics for each Sentiment Class

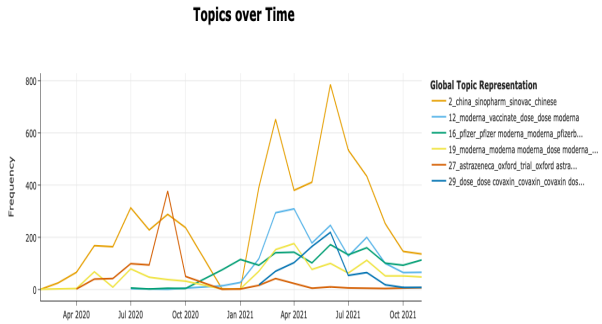

Figure 7: Topics Trends Over Time

Figure 8: Sample Named Entity Recognition for a tweet

	PERSON	NORP	ORG	GPE
Negative	Bill Gate	Chinese	Moderna	China
Positive	Anthony Fauci	Indian	Oxford	China, India
Neutral	Mark Zuckerberg	African	Sinopharm	Canada

Figure 9: Named Entity Recognition

started to wane. Notably, the data indicates that the Sinopharm vaccine garnered twice as much attention in online discussions compared to other vaccines.

Figure 8 provides a visual representation of the different entities that can be detected in a tweet. However, our research focuses specifically on the entities categorized as PERSON, NORP, ORG, and GPE, as these are most pertinent for our investigation into the famous individuals, nationalities, and organizations involved in the online discourse. The count of each entity type identified within the top 10 dominant topics is detailed in Table 1. Figure 9 depicts the dominant entity identified in each sentiment for desired entity type.

5 Conclusion

This study utilised social media analytics to gain insight into the public sentiment and perceptions towards the COVID-19 vaccine. Although the results provided valuable information, it is important to acknowledge the limitations of the study. One of the main limitations is the potential bias in data collection, as the data only includes Twitter users who actively express their views on the vaccine. Additionally, the VADER sentiment analysis tool may not be reliable in capturing the sentiment of tweets with sarcasm or irony. Moreover, in this study we implemented RoBERTa to generate labels for a subset of the dataset. Generating sentiments for the entire dataset required extensive computa-

TOPIC	PERSON	NORP	ORG	GPE
china_sinopharm_sinovac_chinese	9	0	1	0
covaxin_covishield_india_dose	1989	2942	1453	5552
sputnikv_sputnik_russia_russian	1283	147	710	238
vaxxers_vaxxer_antivaxxer_vaxers	875	1700	1485	3547
microchip_chip_track_gate	938	696	1300	2765
pfizer_pfizerbiotech_biotech_pfizer...	904	173	1261	321
moderna_dose moderna_moderna moderna	702	600	1037	246
covaxin_dose_dose covaxin_covaxin covashield	287	95	313	77
ocgn_ocugen_ocgn covaxin_covaxin ocgn	533	58	430	346
astrazeneca_oxford_trail_oxford astra...	582	56	442	403

Table 1: Count of Each Type of Entities in 10 top Dominant Topics

tional resources, which resulted in the exclusion of this methodology from the study. Furthermore, the BERTopic model used in this study may not be able to identify more nuanced topics or themes related to the COVID-19 vaccine, and the spaCy NER model may not accurately recognize all entities in the text.

To improve future research, more advanced natural language processing techniques such as deep learning algorithms can be incorporated to enhance the accuracy of sentiment analysis. Furthermore, more sophisticated topic modeling techniques can be used to capture more complex topics and themes. Additionally, utilising advanced models and techniques to improve entity recognition can help better identify misspelled or unconventional language. In conclusion, social media analytics can provide valuable insights into public perception and sentiment towards the COVID-19 vaccine, but it is important to consider the limitations and continue to improve the methods used in such studies.

References

[1] D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," *IEEE Intelligent Systems*, vol. 25, no. 6, pp. 13–16, 2010.

[2] K. Kurniawati, G. Shanks, and N. Bekmamedova, "The business impact of social media analytics."

[3] H. Xue, X. Gong, and H. Stevens, "Covid-19 vaccine fact-checking posts on facebook: Observational study," *Journal of Medical Internet Research*, vol. 24, no. 6, p. e38423, 2022.

[4] A. Chinnov, P. Kerschke, C. Meske, S. Stieglitz, and H. Trautmann, "An overview of topic discovery in twitter communication through social media analytics."

[5] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351.

[6] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, pp. 617–663, 2019.

[7] X. Zhou, X. Tao, M. M. Rahman, and J. Zhang, "Coupling topic modelling in opinion mining for social media analysis," in *Proceedings of the international conference on web intelligence*, pp. 533–540.

[8] Y. Nie, Y. Tian, X. Wan, Y. Song, and B. Dai, "Named entity recognition for social media texts with semantic augmentation," *arXiv preprint arXiv:2010.15458*.

[9] A. Ritter, S. Clark, O. Etzioni *et al.*, "Named entity recognition in tweets: an experimental study," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1524–1534.

[10] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," in *Proceedings of the international AAAI conference on web and social media*, vol. 7, no. 1, pp. 400–408.

[11] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*.

[12] G.-A. Asderis, "Sentiment analysis on twitter data, a detailed comparison of textblob and vader," 2022.