

## Sentiment Analysis of Covid Vaccine Myths using Various Data Visualization Tools

Tarandeep Kaur Bhatia<sup>1\*</sup>, Samagya Rathi<sup>2</sup>, Thipendra P Singh<sup>3</sup>, and Biswayan Naha<sup>4</sup>

<sup>1,2,4</sup>School of Computer Science, University of Petroleum and Energy Studies (UPES), Dehradun, Uttarakhand, India

<sup>3</sup>School of Computer Science Engineering & Technology, Bennett University, Greater Noida, NCR, India

### Abstract

**INTRODUCTION:** Anti-vaccination agitation is on the rise, both in-person and online, notably on social media. The Internet has become the principal source of health-related information and vaccines for an increasing number of individuals. This is worrisome since, on social media, any comment, whether from a medical practitioner or a layperson, has the same weight. As a result, low-quality data may have a growing influence on vaccination decisions for children.

**OBJECTIVES:** This paper will evaluate the scale and type of vaccine-related disinformation, the main purpose was to discover what caused vaccine fear and anti-vaccination attitudes among social media users.

**METHODS:** The vaccination-related data used in this paper was gathered from Reddit, an information-sharing social media network with about 430 million members, to examine popular attitudes toward the vaccine. The materials were then pre-processed. External links, punctuation, and bracketed information were the first things to go. All text was also converted to lowercase. This was followed by a check for missing data. This paper is novel and different as Matplotlib, pandas, and word cloud was used to create word clouds and every result has a visual representation. The Sentiment analysis was conducted using the NLTK library as well as polarity and subjectivity graphs were generated.

**RESULTS:** It was discovered that the majority population had neutral sentiments regarding vaccination. Data visualization methods such as bar charts showed that neutral sentiment outnumbers both positive and negative sentiment.

**CONCLUSION:** Prevalent Sentiment has a big influence on how people react to the media and what they say, especially as people utilize social media platforms more and more. Slight disinformation and/or indoctrination can quickly turn a neutral opinion into a negative one.

**Keywords:** Vaccine myths, Sentiment Analysis, Reddit, word cloud, social media analysis

Received on 27 December 2023, accepted on 28 March 2024, published on 04 April 2024

Copyright © 2024 T. K. Bhatia *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](#), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetpht.10.5639

\*Corresponding author. Email: [drtarandeepkaurbhatia@gmail.com](mailto:drtarandeepkaurbhatia@gmail.com)

### 1. Introduction

Vaccinations are frequently referred to as the biggest success in the health of the public in contemporary medical history. This is because of immunizations' significant influence on the lowering of infectious illness occurrence [1,2]. Vaccines considerably lowered the financial burden of infectious illness while also improving people's quality of life. Vaccination policies differ from country to country. In 1892, India implemented the Compulsory Vaccination Act to increase smallpox vaccination coverage and lessen the pandemic. Except during epidemics, the 'Act' mainly stayed on the paper [3].

According to sources, the rule was in effect in nearly 80% of British India's districts in 1938. In India, the Expanded Programme of Immunization (EPI) was launched in 1978. (EPI). After gaining traction in 1985, the initiative was extended into the Universal Immunization Program (UIP), which was to be phased in to cover all country's districts by 1989-90. The anti-immunization campaign, on the other hand, is growing in popularity both internationally. and in India, especially after the global Covid-19 Pandemic [4]. Vaccines have been met with skepticism and a wide range of critical viewpoints and behaviours, such as vaccination refusals or aggressive anti-vaccination movements, which have the ability to undo the public health benefits of vaccination. Vaccine aversion is

considered a reluctance in taking or denying immunization even when the service is readily available [5]. Vaccine hesitation was documented when vaccinations were first introduced, in the 18th century when the smallpox vaccine was developed after the smallpox outbreak. "Anti-vax" or "anti-vaxxers" are those who are absolutely opposed to vaccination. Not all vaccination skeptics, however, are "anti-vaxxers." Vaccine reluctance is actually a continuum that spans total acceptance to extreme rejection, with uncertain people sitting somewhere in between. Vaccine-averse people may take some immunizations but reject, postpone, or have reservations about others. This misunderstanding highlights the necessity for a thorough knowledge of the elements that lead to vaccine hesitancy development [6].

Anti-vaccination activism is growing both live and online these days, particularly on social media. For a growing number of individuals, the World Wide Web is now a primary source of health-related information and immunizations. This is concerning, since, on social media, each statement has a similar impact, regardless of whether it comes from a medical professional or a layperson. As a result, low-quality information may increasingly impact decision-making in the area of children's vaccination [7]. As a result, there is growing worried over antivaccine sentiment in many countries, prompting rapid advancements in vaccination hesitancy research. The usefulness of such research resides in the information they provide about which content concepts elicit the greatest apprehension and negative feelings, as well as those that are more receptive to modification, allowing for more precise and successful targeting of online information campaigns [8]. It's crucial to know ahead of time what the community thinks about the vaccine policy's implementation. This research may also help healthcare workers connect more effectively with reluctant patients who may be convinced to be vaccinated. People may not only have reservations about the efficacy or safety of vaccines, according to studies but may also trust numerous conspiracy ideas or feel that other medicines might offer more disease protection than immunizations [7, 8]. Preparing for and responding to such arguments, even if only obliquely, may improve the efficacy of reassurance and persuasion. Sentiment analysis is one tool that may be used to better understand public attitudes about vaccination. It reveals which emotions are most prevalent in vaccination-related utterances, allowing us to figure out which way of distributing vaccine information and reassuring apprehensive parents would be most economical [9]. This pragmatic approach towards immunization information treats vaccines like a product for sale, provides workers in the field of public health with useful tools to manage vaccine apprehension, and has been used in various studies on the subject. Sentiment analysis extracts and classifies sentiments using Natural Language Processing (NLP), a text analysis and computing approach.

## 2. Background

Sentiment analysis is the classification of written texts into positive, negative, or neutral polarity values to determine the sentiment of a subject, concept, phenomenon, or event [10]. These approaches have already been employed in numerous scientific, social, and economic applications because assessing public emotion is critical for selecting suitable messaging, interventions, and policy. Social media sentiment analysis is a relatively young discipline. Data from social media is being utilized in a wide range of research projects. People have wondered if microblogs (like Twitter) are better for sentiment analysis than larger publications. Another research looked at the 2012 U.S. Presidential election using sentiment analysis tools. Despite numerous concerns about social media data's validity, representativeness, confounding, and biases with close to 390 crore members globally, social networking sites continue to be a valuable source of textual, semantically rich data that may be used to track a wide range of social interactions, including talks about public health issues [11].

Many initiatives to monitor posts on social networking sites about the occurrence of vaccination and illness have been quite effective credit to Natural Language Processing (NLP) tools. Alessa and Miad (2019) tracked influenza-related tweets on Twitter and recognized the beginning of a surge in cases [12]. Raghupathi et al. (2020) discovered a correlation between positive attitudes and successful public health strategies, and between a surge in cases of measles and unfavourable negative statements on the measles vaccine. These technologies are now used to observe popular perceptions of the use of masks during an outbreak across the world. To assess public opinion on a number of COVID-19-related concerns, several researchers employed a combination of machine learning classification algorithms and sentiment analysis approaches. By early 2021, various sentiment analysis research on the coronavirus vaccination were already conducted. Gbashi et al. (2021) utilized news articles to detect the view of media polarity on the coronavirus vaccination in Africa. In addition, the study further examined public opinion in India, Indonesia, and China. Research on the public perception of the coronavirus vaccine and topic modelling of various subreddits was presented by Wu et al., 2021 [13]. The subreddits selected for this study, on the other hand, contained discussions on the novel coronavirus vaccine which were not focused or centered on the vaccine in a direct sense.

## 3. Literature Review

This section discusses a diverse literature survey of approximately 20 research papers. Although this paper is prepared after reviewing around 35 research and review papers.

- 1) A deep investigation of the coronavirus vaccination talks on Twitter has been undertaken in this study. From the standpoint of nations and vaccination brands, the study looked at the hot themes addressed by individuals and related emotional polarity [14]. The findings revealed that the majority of individuals believe vaccinations work and are anxious to be vaccinated. Negative tweets, on the other hand, were frequently linked to news stories about mortality followed by vaccination, shortage of vaccines, and side effects of injections. As a whole, the researchers employed common Natural Language Processing (NLP) technology to collect and objectively assess and depict the thoughts of people on the coronavirus vaccination on social media. Their discoveries increased the readability of ambiguous materials on social media platforms and provided the administration with useful data support.
- 2) In this paper the Twitter API was used to collect the Covid-19 vaccine topic. The VADER model was then implemented to assess the sentiment groups (positive, neutral, and negative) and determine the sentiment scores of the sample using unsupervised learning. The Latent Dirichlet Allocation (LDA) model was applied to take out themes and main words after determining the number of topics. People's feelings about the Chinese vaccination differed from those about vaccines in other nations, and the sentiment score may have been influenced by the count of every day, new cases, and fatalities, as well as the type of major concerns in the communication network [15].
- 3) This study aimed to evaluate the emotions of people during the outbreak using sentiment analysis and natural language processing techniques to recognize texts and extract polarity, emotion, or unanimity of coronavirus vaccines on the basis of tweets. The method used a collection of tweets, with the text parsed using the nltk toolkit and the keywords generated by the Tf-IDF algorithm [16]. It counted how many n-gram keywords and hashtags were used in tweets. The findings revealed that emotions are separated into two categories: good and negative, with negative emotions prevailing.
- 4) Using the LDA method, the researchers attempted to confirm the COVID-19-related vaccination myths and rumours in this work. For the experiment, they employed data mining, text mining, and sentiment analysis. The experiment's findings revealed that while most individuals are pleased about vaccination, it also has a negative influence. The majority of individuals, according to the experiment, are talking about "vaccines," "people," "morons," and "ever." They provided a method for validating vaccination myths of this nature. When compared to other frameworks, LDA algorithms were able to anticipate and confirm the myth up to 70% of the time. Their experimental outcome demonstrates promising efficiency [17].
- 5) This study aimed to evaluate and analyze papers over the previous 11 years that dealt with various vaccine hesitancy situations to better understand how sentiment analysis might be applied to suitable literature results. Papers and pieces were carefully searched on various publication sites. 30 articles were picked based on the criteria of inclusion and exclusion. These papers were categorized into a literary taxonomy, which included problems, motives, and suggestions for public health, medical, social, and technical disciplines. Insightful patterns were discovered, and chances for better knowledge of this phenomenon were fostered [18].
- 6) The purpose of this project was to construct a system capable of automatically detecting people that propagate anti-vaccine narratives to understand anti-vaccine emotions in a better way and strive to limit its influence. Researchers released a freely accessible Python script that analyses Twitter accounts to determine the likelihood of spreading antivaccine sentiment in the future. Automated dataset generation, neural networks, and text embedding methods were used to create the software package. It has been trained on over 100,000 accounts and millions of tweets [19]. This approach assists academics and policymakers in comprehending antivaccine conversation and disinformation methods, which can then be used to build focused campaigns aimed at informing and debunking the harmful anti-vaccination beliefs presently circulating.
- 7) The goal of this work was to focus on the latest improvements in transposer-based machine learning approaches and see if transformer-based machine learning can be utilized to analyse the attitude toward vaccination during pregnancy represented in social media posts. Using keyword searches relevant to maternal vaccination, 16,604 tweets were selected. After removing extraneous tweets, the remaining tweets were categorized into three groups: promotional, discouraging, ambiguous, and neutral by three different researchers. Multiple machine learning approaches were applied to a portion of the final data set of 2722 unique tweets, then evaluated and compared by human analysts. When the aggregated score of the three annotators was compared, the machine learning algorithms were determined to be 81.8 percent

accurate (F score=0.78). Individual annotators' accuracies relative to the final score were 83.3 percent, 77.9%, and 77.5 percent, respectively. The study found that by utilizing machine learning models, we can reach very near the accuracy in grouping tweets as one human coder can. The ability to employ this automated, dependable, and precise procedure might save time and resources for completing this analysis, as well as guiding potentially useful and essential solutions [20].

- 8) The focus of this research was to find the views and emotions of people in coronavirus vaccine-related social media discussions and to recognize significant changes in perceptions over a period to understand public views, queries, and feelings that might affect herd immunity aspirations. Tweets were retrieved from a large-scale coronavirus data collection. The Tweets were cleaned with R software and those that had the phrases vaccinate, vaccine, immunization, vaccination, vaccines, and vaccinations were kept. The total number of tweets considered in the research was 1,499,421 from 583,499 distinct people. Using the Latent Dirichlet Allocation for topic modelling, sentiment and emotion analysis using R was performed. Topic modelling of corona vaccine-related tweets revealed 16 subjects that were organized in five overarching themes. The most tweeted most issue (227,840/1,499,421 tweets, 15.2 percent) was vaccination opinions, which remained a hot topic throughout the time we studied it [21]. Around August of 2020, when Putin announced that Russia has developed the first coronavirus vaccine, vaccine development became a hot topic of discussion. With the improvement of vaccine delivery, the issue of vaccination teaching grew increasingly important, eventually becoming the most talked-about topic at the beginning of January 2021. Despite the fluctuations, weekly average sentiment values depicted that sentiment was generally improving.
- 9) This study used Twitter as the raw data to examine various research on sentiment analysis of the coronavirus vaccination. It was carried out to learn how researchers gather, preprocess, and categorize data [22]. To gather, filter, and review the research papers discovered, a systematic literature review technique was employed. To further investigate the material of the articles, they were scanned on the basis of exclusion and inclusion criteria. As a consequence of this approach, several investigations were eliminated, the remaining twenty-one relevant papers were investigated for this study. The remaining articles were then reviewed to find answers to the four research questions. Twitter, API, Rtweet,

Tweepy, and Twint are some of the most often utilized data collection tools. Lemmatization, stop word removal, removing duplicate tweets, removal of punctuation and link stemming, and tokenization were some of the approaches used to preprocess Twitter data.

- 10) The exploratory data analysis phase of this project's major purpose is to become familiar with the columns and begin developing research topics, which will be projected utilizing data visualization to examine the data ecosystem [23]. This study examined public reaction to the immunization campaign using Twitter data acquired between February 3rd and February 10th, 2021. A dataset to examine the impact of Tweet sentiment on Twitter users was utilized for this study. This data categorized 2,000 separate Tweets by favorite count, location, description, number of followers, emotion, and other factors. A lot of extremely subjective sentiments when we look at the top 10 most positively charged tweets. "Hope," "gratitude," "safety," "support," and "joy" are among the most often used words and phrases. Another finding from this study is the correlation between tweet polarity and the number of favorites. People like tweets that are impartial or non-polarized.
- 11) The goal of this paper was using social sites to track down and study vaccine opponents' grounds against vaccines in childhood. All remarks made by people on the Facebook page of a famous Polish opponent were gathered and quantitatively assessed in terms of what they contained using Kata's modified technique [24]. There was also sentiment analysis. The 4,042 comments that met the criteria included conspiracy ideas (28.2%), misleading information and unverified establishment (19.9%), failure to comply with civil liberties (13.2%), content pertaining to the safety and efficiency of vaccinations (14.0%), personal observations (10.9%), moral standards, faith, and assumption (8.5%), and alternative therapies (8.5%). 1,223 were comments in favour of vaccines, 15.2% of which were hostile, insulting, or nonsensical. Opinions without any justification, and those including claims about other medicinal options or disinformation, appeared to be more optimistic and less furious than those grouped in some other topics, according to sentiment analysis. A considerable proportion of materials in conspiracy theory and disinformation groups suggested that the writers of remarks like these lack faith in medicine's scientific achievements. The paper concluded that vaccination efforts should fully address these findings.

- 12) The researchers utilized a hybrid technique to analyse 1,499,227 tweets related to vaccines out of which 69.36% were classified as neutral, 8.86 percent as negative, and 21.78 percent as positive, by the system. The proportion of neutral tweets decreased with time, but the ratio of positive and negative tweets grew [25]. Every April, there were peaks observed in favourable tweets. The percentage of favourable tweets was much greater during the week and significantly lower on weekends. The pattern for negative tweets was the polar opposite. 91.83 percent of people with two tweets had a homogenous polarized dialogue. In Switzerland, positive tweets were more common (71.43 percent), Netherlands (15.53 percent), Canada (11.32 percent), Japan (10.74 percent), and US were the countries with the most negative tweets (10.49 percent). The study concluded that opinion mining might be useful for tracking internet vaccination-related concerns and adopting vaccine awareness methods as required.
- 13) 9581 tweets related to vaccination were collected for this study. The sample is subjected to sentiment analysis, with the data being clustered into themes using the TF-IDF approach. According to the findings, most of the tweets (approx. 77%) were about the hunt for new/better vaccinations for illnesses including Ebola, HPV, and flu. About half of the others were concerned about the measles outbreak in the US, and the remaining were involved in current disputes between advocates and adversaries of the measles vaccine [26]. While these figures imply that vaccination misinformation plays a little influence, the idea of herd immunity puts that role into context. The study concluded that health specialists should anticipate the possibility of lies being increasingly widespread in the public psyche in the future.
- 14) The name/headlines of videos released on YouTube throughout these times were used in this article trying to assess if and how the public's perception changed before and after the vaccine campaign [27]. The researchers investigated the subjects of YouTube Italian videos on vaccinations in 2017 and 2018 using sentiment analysis. Vaccinations were strongly opposed before the law, according to the CON. Instead, individuals become less critical following the PR effort. According to sentiment research, the aggressive vaccination campaign, which was also advocated by physicians, drove the attitude to shift from a polarized negative view in 2017 (52 percent negative) to a polarized good opinion in 2018. (54 percent positive). The research suggested that social media vaccination campaigns might be an important tool for health policymakers and a powerful weapon against ignorance and misinformation from unqualified persons affecting people's decisions.
- 15) The goal of this paper was to perform sentiment analysis on the topic of coronavirus immunization, perform temporal and spatial analyses of the textual data, and identify the most often debated themes so organizations can raise awareness about them [28]. The study used 14 different machine learning classifiers and natural language processing to analyse the sentiment of tweets (NLP). TextBlob and Vader, both based on a lexicon, are used to annotate the data. Textual data were pre-processed using a natural language toolset. For all four datasets, the researchers found that the unigram model performed better than bigram and trigram models. Models that used the TF-IDF were on average more precise than the count vectorizer models. The mean precision of logistic regression in the count vectorizer class was 91.925 percent. All unigram models had a standard deviation  $< 1$ , except for the Gaussian Nave Bayes, which had an SD of 1.18. The findings of the experiment demonstrated the days and hours when the happiest, negative, and neutral tweets were sent out.
- 16) The purpose of this paper was to ensure a new coronavirus vaccine disinformation detection system based on machine learning. The researchers used machine learning techniques to classify vaccination misinformation after collecting and annotating tweets on COVID-19 vaccine [29]. Using credible sources and health specialists, over 15,000 tweets were marked as misleading or generic vaccination tweets. The highest accuracy result was obtained using BERT, which yielded a 0.98 F1 score on the test set. The accuracy and recall scores were, respectively, 0.97 and 0.98. The study found that machine learning-based algorithms are successful in detecting coronavirus vaccine misinformation on social networking sites.
- 17) The researchers performed a survey in the US to gain knowledge regarding individuals' health literacy about vaccination as well as their COVID-19 views and experiences [30]. For the research, a nationwide specimen was taken, with key demographics (ethnicity, age, race, and gender) roughly matching the United States Census percentages. IBM SPSS version 27 and  $\chi^2$  tests were used to evaluate the data, with Z-tests used to establish more detailed comparisons across groups. Individuals who believed the coronavirus vaccine to be harmful were less likely to accept it, those who knew little regarding the virus were likely to accept myths

about the vaccination. The immunization was deemed safe by less informed, lower-income populations living in rural places. The findings emphasize on the need to clearly communicate health information to people of all incomes and education settings.

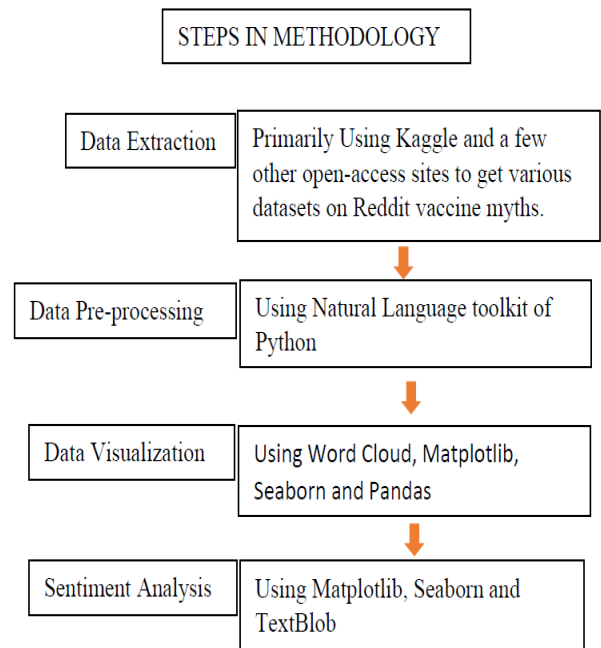
- 18) The researchers attempted to determine the sentiment of a common individual's reaction to the coronavirus vaccination process from the Bengali text using several classification analysis techniques based on machine learning [31]. The researchers gathered information from polls on the internet and several media sites and divided it into three categories: neutral, negative, and positive. The researchers used LSTM and CNN, two deep learning algorithms, Logistics Regression, Nave Bayes, K-nearest neighbors, SVM, Decision Tree, and Random Forest to develop six prominent classification techniques. CNN has the highest accuracy of 65.41 percent among them
- 19) The primary goal of the paper was to utilize data analysis techniques to examine the survey responses and draw conclusions. The researchers performed a Sentimental Analysis of participant replies to determine what prevents people from being immunized [32]. They analyzed and visualized 200 replies to 11 questions in this study. The survey included data on age, gender, area, vaccination willingness, preexisting medical issues, and vaccine choice. Doctors and the general public both filled out the form. Graphs, charts, sentiment analysis, word clouds, and other tools were used to examine the data. The majority of the individuals preferred vaccinations developed in India, according to the researchers. Participants who did not have any preexisting underlying problems have also experienced vaccination adverse effects. They also discovered that while many people are fearful of the vaccine's adverse effects, the majority of people are anxious to get vaccinated.
- 20) Latent Dirichlet Allocation, topic modelling, and sentiment analysis were performed on data in the form of text obtained from 13 Reddit groups centered on the coronavirus vaccine, to examine the immunization-related debate in social media. To discover changes in sentiment and latent themes, data was aggregated and examined for a month [33]. These groups indicated more positive sentiment than negative sentiment towards vaccine-related topics, according to the polarity analysis, which has remained constant throughout time. Topic modelling suggested that community members were more concerned with adverse effects than with bizarre conspiracy ideas. The emotions stated in such types of

individuals are more favourable than negative and have not altered much since December of 2020.

## 4. Methodology

### *Methodology and Tools Used*

The methodology used in the paper is illustrated in Figure 1. Python Coding was done in Google Colab Notebook. Necessary Python libraries and modules, namely, NumPy, pandas, matplotlib, seaborn, Word Cloud, nltk, and text Blob were imported and used to generate the required word clouds and graphs [34].



**Figure 1.** Methodology used

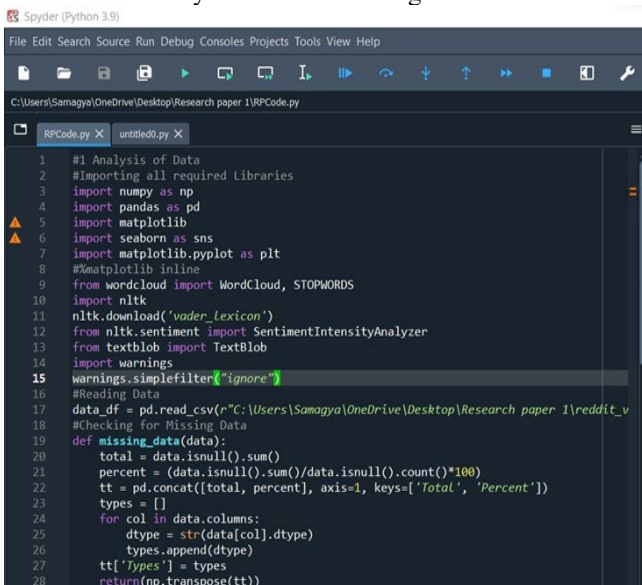
### *Data Source*

The vaccination-related data was gathered from Reddit, an information-sharing social media network with about 430 million members, to examine popular attitudes toward the vaccine. The platform is made up of user-created groups (subreddits), each with its own set of community rules. Members of the subreddit can submit links, photographs, videos, and text. People in the community then often upvote or downvote a post on the basis of the judgment of its quality and write comments. Posts then are categorized as hot, new, rising, or controversial based on the distribution of votes. Following this, the most loved posts in every category are promoted to the top of the community's page [35]. The same vote ranking method applies to these comments. The upvote/downvote method on Reddit is designed to improve the quality of articles by removing irrelevant content. The data was collected from various open-access datasets on the internet. Then, the removal of all irrelevant information was done and combined these datasets into

one CSV file. After the data was cleaned, structured, and organized, we proceeded with the sentiment analysis.

### Setting up the Environment & Data Pre-processing

Imported all necessary python libraries and modules, namely, NumPy, pandas, matplotlib, seaborn, word cloud, nltk, and text blob as shown in Figure 2. For effective model training, the contents must be pre-processed. External links, punctuation, and bracketed content were deleted first [36]. All text was changed to lowercase as well. Stop words are often used in terms like 'the,' 'and', 'in,' and 'for.' The difficulty of training can be reduced by doing away with low-information terms that give minimal contextual information. The NLTK30 Python package was used to complete this stage. Stemming, a preprocessing step, lowers the number of derivationally related forms in a word to a single base form as elaborated in Figure 3. The NLTK library was utilized in this stage. This was followed by a check for missing data.

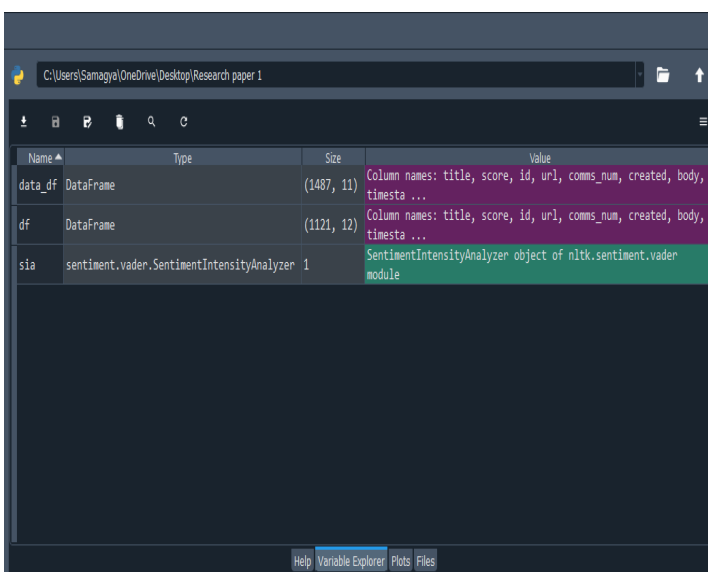


```

1 #1 Analysis of Data
2 #Importing all required Libraries
3 import numpy as np
4 import pandas as pd
5 import matplotlib
6 import seaborn as sns
7 import matplotlib.pyplot as plt
8 #%matplotlib inline
9 from wordcloud import WordCloud, STOPWORDS
10 import nltk
11 nltk.download('vader_lexicon')
12 from nltk.sentiment import SentimentIntensityAnalyzer
13 from textblob import TextBlob
14 import warnings
15 warnings.simplefilter("ignore")
16 #Reading Data
17 data_df = pd.read_csv(r"C:\Users\Samagya\OneDrive\Desktop\Research paper 1\reddit_v
18 #Checking for Missing Data
19 def missing_data(data):
20     total = data.isnull().sum()
21     percent = (data.isnull().sum()/data.isnull().count()*100)
22     tt = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
23     types = []
24     for col in data.columns:
25         dtype = str(data[col].dtype)
26         types.append(dtype)
27     tt['Types'] = types
28     return(np.transpose(tt))

```

Figure 2. Setting up the Environment

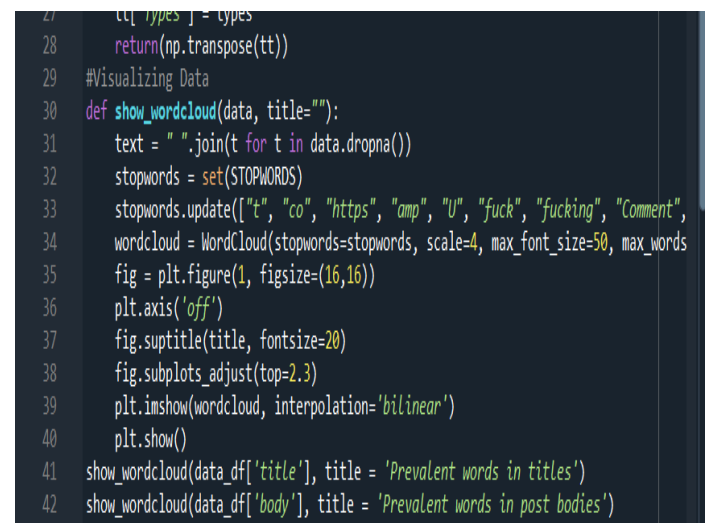


Name	Type	Size	Value
data_df	DataFrame	(1487, 11)	Column names: title, score, id, url, comms_num, created, body, timesta...
df	DataFrame	(1121, 12)	Column names: title, score, id, url, comms_num, created, body, timesta...
sia	sentiment.vader.SentimentIntensityAnalyzer	1	SentimentIntensityAnalyzer object of nltk.sentiment.vader module

Figure 3. Data Pre-processing

### Data Visualization

Word clouds are amazing at representing the most frequent words in titles and the body of the post [37]. The magnitude of each word in a word cloud, which is a data visualization tool for viewing text data, shows its frequency or significance as elaborated in Figure 4. A word cloud is employed to emphasize important textual information. Word clouds are widely used to evaluate data from social networking websites as displayed in Figure 5 and Figure 6. Matplotlib, pandas, and word Cloud are the modules used to create the word cloud. Unnecessary frequent words like 'vaccine', 'vaccination', 'and', 'comments' are removed. The title is determined by keyword comments.



```

27 cc['types'] = types
28 return(np.transpose(tt))
29 #Visualizing Data
30 def show_wordcloud(data, title=""):
31     text = " ".join(t for t in data.dropna())
32     stopwords = set(STOPWORDS)
33     stopwords.update(["t", "co", "https", "amp", "u", "fuck", "fucking", "Comment",
34     wordcloud = WordCloud(stopwords=stopwords, scale=4, max_font_size=50, max_words
35     fig = plt.figure(1, figsize=(16,16))
36     plt.axis('off')
37     fig.suptitle(title, fontsize=20)
38     fig.subplots_adjust(top=2.3)
39     plt.imshow(wordcloud, interpolation='bilinear')
40     plt.show()
41 show_wordcloud(data_df['title'], title = 'Prevalent words in titles')
42 show_wordcloud(data_df['body'], title = 'Prevalent words in post bodies')

```

Figure 4. Visualization of Data

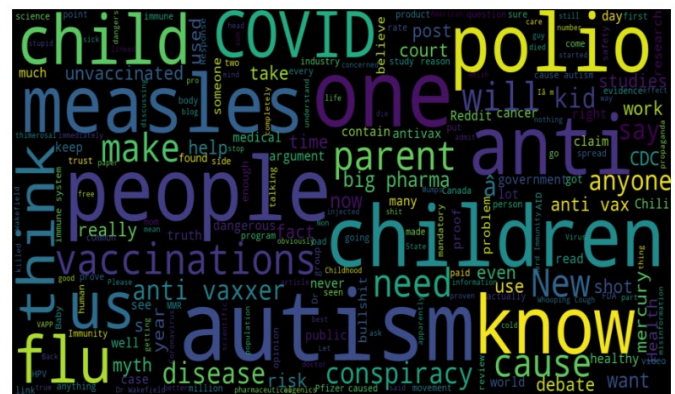
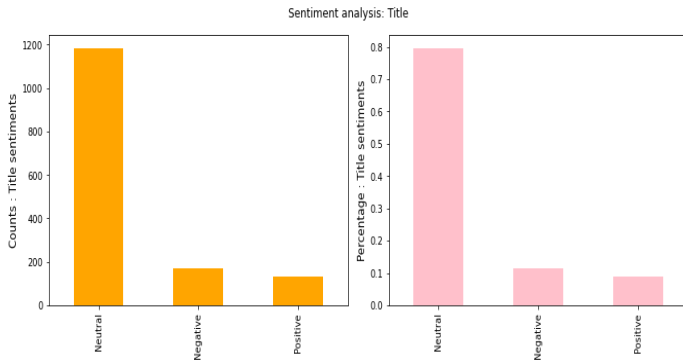


Figure 5. Prevalent words in Titles

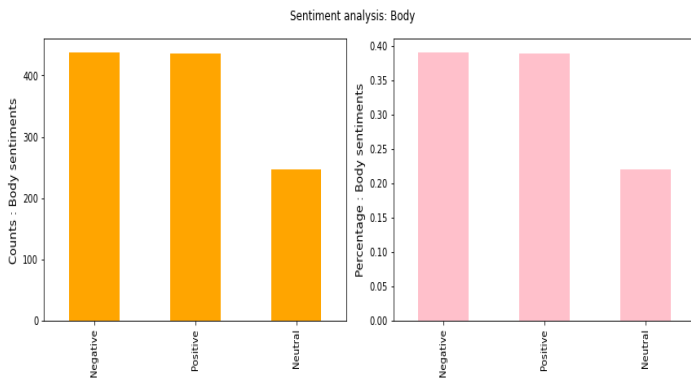




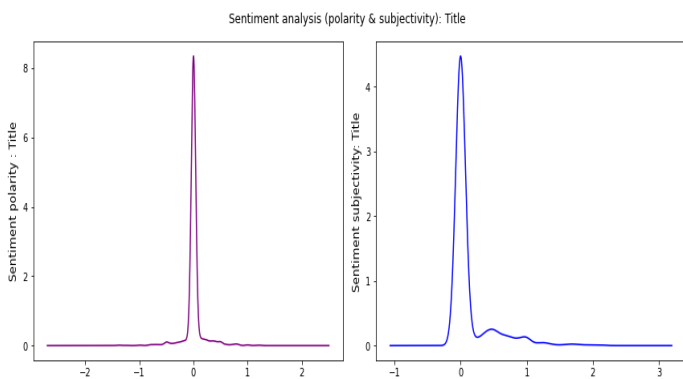




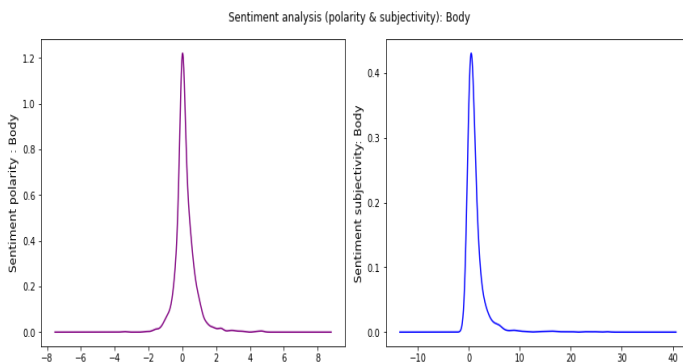
**Figure 16.** Sentiment Analysis of prevalent words in the title



**Figure 17.** Sentiment Analysis of prevalent words in the body



**Figure 18.** Polarity and Subjectivity for Title



**Figure 19.** Polarity and Subjectivity of Body

## 6. Discussion

Reddit data to examine public opinion on vaccinations was used for this study, which has a significant impact on how society reacts to the media, and what they might say, especially as the use of social media platforms grows every day. A Neutral opinion might be easily manipulated into a negative one by slight misinformation and/or brainwashing. This may in turn develop vaccine hesitancy in a large population. Thus, the concerned authorities need to ensure that they contain the spread of misinformation and fake news. Although currently, society at large is willing to get vaccinated provided, the procedure is hassle-free and easy. The vaccination speed can be boosted by encouraging the general public and educating them about the benefits of the vaccine. Given the Covid-19 pandemic, such steps should be taken in an efficient and timely fashion to achieve herd immunity and stop this notorious virus from creating even greater chaos.

## 7. Open Research Challenges And Future Trends

This study has only considered the data from one social media, that is, Reddit. The data is limited to certain regions and thus limits us from forming a larger picture. Sentiment Analysis of people on social media also ignores the population which is not so present online. In certain countries where the Internet is not very accessible or internet freedom is limited, this analysis will prove to be futile because of the very limited sample. In many cases, the views of a person on social media might be misleading and differ from one platform to another. Thus, it is up to further researchers to determine how the opinions of people having a neutral opinion can be easily made extreme to either end and in what way. Research is also needed to determine if the social media sentiment analysis is in line with vaccination rates on the ground. Profiling individuals based on the various social media apps they use and then predicting if a particular person will get vaccinated or not can also be a topic for further research and analysis. How taking a vaccine affects sentiment is also left for future studies.

## 8. Conclusion

For this study, Reddit data to look at public opinion on vaccines was utilized, which has a big influence on how people respond to the media and what they say, especially since the usage of social media platforms expands every day. Neutral sentiment tends to outnumber both positive sentiment and negative sentiment, as seen by data visualization tools like bar chart. Currently, the general public is eager to get vaccinated if the treatment is simple and painless. The responsible authorities must ensure that disinformation and fake news are not circulated. The rate of vaccination can be accelerated by promoting and

educating the general population about the vaccine's advantages.

## References

- [1] Ansari, Md Tarique Jamal, and Naseem Ahmad Khan. "Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content." *Electronic Journal of General Medicine* 18, no. 6 (2021).
- [2] Yin, Hui, Xiangyu Song, Shuiqiao Yang, and Jianxin Li. "Sentiment analysis and topic modeling for COVID-19 vaccine discussions." *World Wide Web* 25, no. 3 (2022): 1067-1083.
- [3] Mohan, Sumit, Anil Kumar Solanki, Harish Kumar Taluja, and Anuj Singh. "Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach." *Computers in Biology and Medicine* 144 (2022): 105354.
- [4] Ali, GG Md Nawaz, Md Mokhlesur Rahman, Md Amjad Hossain, Md Shahinoor Rahman, Kamal Chandra Paul, Jean-Claude Thill, and Jim Samuel. "Public perceptions of COVID-19 vaccines: Policy implications from US spatiotemporal sentiment analytics." In *Healthcare*, vol. 9, no. 9, p. 1110. MDPI, 2021.
- [5] Hananto, Andhika Rafi, Silvia Anggun Rahayu, and Taqwa Hariguna. "COVID-19 Vaccination: A Retrospective Observation and Sentiment Analysis of the Twitter Social Media Platform in Indonesia." *International Journal of Informatics and Information Systems* 5, no. 1 (2022): 56-68.
- [6] Ali, GG Md Nawaz, Md Mokhlesur Rahman, Md Amjad Hossain, Md Shahinoor Rahman, Kamal Chandra Paul, Jean-Claude Thill, and Jim Samuel. "Public perceptions of COVID-19 vaccines: Policy implications from US spatiotemporal sentiment analytics." In *Healthcare*, vol. 9, no. 9, p. 1110. MDPI, 2021.
- [7] Yadav, Ashima, and Dinesh Kumar Vishwakarma. "A Language-independent Network to Analyze the Impact of COVID-19 on the World via Sentiment Analysis." *ACM Transactions on Internet Technology (TOIT)* 22, no. 1 (2021): 1-30.
- [8] Riyanto, Riyanto, and Abdul Azis. "Application of the Vector Machine Support Method in Twitter Social Media Sentiment Analysis Regarding the Covid-19 Vaccine Issue in Indonesia." *Journal of Applied Data Sciences* 2, no. 3 (2021): 102-108.
- [9] Kwok, Stephen Wai Hang, Sai Kumar Vadde, and Guanjin Wang. "Tweet topics and sentiments relating to COVID-19 vaccination among Australian Twitter users: machine learning analysis." *Journal of medical Internet research* 23, no. 5 (2021): e26953.
- [10] Shofiya, Carol, and Samina Abidi. "Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data." *International Journal of Environmental Research and Public Health* 18, no. 11 (2021): 5993.
- [11] Eachempati, Prajwal, Praveen Ranjan Srivastava, and Prabin Kumar Panigrahi. "Sentiment analysis of COVID-19 pandemic on the stock market." *American Business Review* 24, no. 1 (2021): 8.
- [12] Mushtaq, Muhammad Faheem, Mian Muhammad Sadiq Fareed, Mubarak Almutairi, Saleem Ullah, Gulnaz Ahmed, and Kashif Munir. "Analyses of Public Attention and Sentiments towards Different COVID-19 Vaccines Using Data Mining Techniques." *Vaccines* 10, no. 5 (2022): 661.
- [13] Alabrah, Amerah, Husam M. Alawadh, Ofonime Dominic Okon, Talha Meraj, and Hafiz Tayyab Rauf. "Gulf countries' citizens' acceptance of COVID-19 vaccines—A machine learning approach." *Mathematics* 10, no. 3 (2022): 467.
- [14] Yin, Hui, Xiangyu Song, Shuiqiao Yang, and Jianxin Li. "Sentiment analysis and topic modeling for COVID-19 vaccine discussions." *World Wide Web* 25, no. 3 (2022): 1067-1083.
- [15] Xu, Han, Ruixin Liu, Ziling Luo, and Minghua Xu. "COVID-19 Vaccine Sensing: Sentiment Analysis and Subject Distillation from Twitter Data." *Available at SSRN 4073419*.
- [16] Sarirete, Akila. "Sentiment analysis tracking of COVID-19 vaccine through tweets." *Journal of Ambient Intelligence and Humanized Computing* (2022): 1-9.
- [17] Roy, Tamal Joyti, Md Ashiq Mahmood, and Aninda Mohanta. "An Efficient Approach to Validate COVID-19 Related Vaccine Myths Utilizing LDA Algorithm." (2022).
- [18] Alamoodi, Abdullah Hussein, B. B. Zaidan, Maimonah Al-Masawa, Sahar M. Taresh, Sarah Noman, Ibraheem YY Ahmaro, Salem Garfan et al. "Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy." *Computers in Biology and Medicine* 139 (2021): 104957.
- [19] Schmitz, Matheus, Goran Murić, and Keith Burghardt. "A Python Package to Detect Anti-Vaccine Users on Twitter." *arXiv preprint arXiv:2110.11333* (2021).
- [20] Kummervold, Per E., Sam Martin, Sara Dada, Eliz Kilich, Chermain Denny, Pauline Paterson, and Heidi J. Larson. "Categorizing vaccine confidence with a transformer-based machine learning model: analysis of nuances of vaccine sentiment in Twitter discourse." *JMIR medical informatics* 9, no. 10 (2021): e29584.
- [21] Lyu, Joanne Chen, Eileen Le Han, and Garving K. Luli. "COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis." *Journal of medical Internet research* 23, no. 6 (2021): e24435.
- [22] Agustiningih, Kartikasari Kusuma, Ema Utami, and Hanif Al Fatta. "Sentiment Analysis of COVID-19 Vaccine on Twitter Social Media: Systematic Literature Review." In *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 121-126. IEEE, 2021.
- [23] Lopez Torres, Ivan. "Twitter-Vaccine Sentiment Analysis." *Available at SSRN 3942987* (2021).
- [24] Klimiuk, Krzysztof, Agnieszka Czoska, Karolina Biernacka, and Łukasz Balwicki. "Vaccine misinformation on social media—topic-based content and sentiment analysis of Polish vaccine-deniers' comments on Facebook." *Human Vaccines & Immunotherapeutics* 17, no. 7 (2021): 2026-2035.
- [25] Piedrahita-Valdés, Hilary, Diego Piedrahita-Castillo, Javier Bermejo-Higuera, Patricia Guillem-Saiz, Juan Ramón Bermejo-Higuera, Javier Guillem-Saiz, Juan Antonio Sicilia-Montalvo, and Francisco Machío-Regidor. "Vaccine hesitancy on social media: sentiment analysis from June 2011 to April 2019." *Vaccines* 9, no. 1 (2021): 28.
- [26] Raghupathi, Viju, Jie Ren, and Wullianallur Raghupathi. "Studying public perception about vaccination: A sentiment analysis of tweets." *International journal of environmental research and public health* 17, no. 10 (2020): 3464.

- [27] Porreca, Annamaria, Francesca Scozzari, and Marta Di Nicola. "Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination." *BMC Public Health* 20, no. 1 (2020): 1-9.
- [28] Jayasurya, Gutti Gowri, Sanjay Kumar, Binod Kumar Singh, and Vinay Kumar. "Analysis of Public Sentiment on COVID-19 Vaccination Using Twitter." *IEEE Transactions on Computational Social Systems* (2021).
- [29] Hayawi, Kadhim, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. "ANTI-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection." *Public health* 203 (2022): 23-30.
- [30] Kricorian, Katherine, Rachel Civen, and Ozlem Equils. "COVID-19 vaccine hesitancy: Misinformation and perceptions of vaccine safety." *Human Vaccines & Immunotherapeutics* 18, no. 1 (2022): 1950504.
- [31] Sarker, Puja, and Dibbendu Kumar Sarker. "Sentiment Analysis of General Peoples Reaction About Covid19 Vaccination in Bangladesh Using Machine Learning Algorithm from Bengali Text Dataset." (2021).
- [32] Dumre, Rutvik, Kopal Sharma, and Karthik Konar. "Statistical and Sentimental Analysis on Vaccination against COVID-19 in India." In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pp. 1-6. IEEE, 2021.
- [33] Dumre, Rutvik, Kopal Sharma, and Karthik Konar. "Statistical and Sentimental Analysis on Vaccination against COVID-19 in India." In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pp. 1-6. IEEE, 2021.
- [34] Kartikasari Kusuma Agustini, Ema Utami, Hanif Al Fatta. "Sentiment Analysis of COVID-19 Vaccine on Twitter Social Media: Systematic Literature Review", 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2021
- [35] Britzolakis, Alexandros, Haridimos Kondylakis, and Nikolaos Papadakis. "AthPPA: A Data Visualization Tool for Identifying Political Popularity over Twitter." *Information* 12, no. 8 (2021): 312.
- [36] Krzysztof Klimiuk, Agnieszka Czoska, Karolina Biernacka, Łukasz Balwicki. "Vaccine misinformation on social media – topic-based content and sentiment analysis of Polish vaccine-deniers' comments on Facebook", *Human Vaccines & Immunotherapeutics*, 2021
- [37] Kandasamy, Venkatachalam, Pavel Trojovský, Fadi Al Machot, Kyandoghere Kyamakya, Nebojsa Bacanin, Sameh Askar, and Mohamed Abouhawwash. "Sentimental Analysis of COVID-19 Related Messages in Social Networks by Involving an N-Gram Stacked Autoencoder Integrated in an Ensemble Learning Scheme." *Sensors* 21, no. 22 (2021): 7582.
- [38] Waheeb, Samer Abdulateef, Naseer Ahmed Khan, and Xuequn Shang. "Topic Modeling and Sentiment Analysis of Online Education in the COVID-19 Era Using Social Networks Based Datasets." *Electronics* 11, no. 5 (2022): 715.
- [39] K. Hayawi, S. Shahriar, M.A. Serhani, I. Taleb, S.S. Mathew. "ANTI-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection", *Public Health*, 2022
- [40] Han Xu, Ruixin Liu, Ziling Luo, Minghua Xu, Bang Wang. "COVID-19 Vaccine Sensing: Sentiment Analysis from Twitter Data", 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021