NUML International Journal of Engineering and Computing
Volume: 2 Issue: 1
https://numl.edu.pk/journals/nijec
Print ISSN: 2788-9629
E-ISSN: 2791-3465
DOI: 10.52015/nijec.v2i1.26

NLP-Based Analysis of Public Perception Regarding Covid-19 Vaccines

Neelofar Memon[a,*],,Zafi Sherhan Syed [a] , Komal Memon [a]

[a]*Department of Telecommunication, Mehran University of Engineering & Technology Jamshoro, Pakistan*
(memon.neelo12@gmail.com, zafisherhan.shah@faculty.muet.edu.pk, komal.memon@admin.muet.edu.pk)

Submitted	Revised	Published
7-March-2023	11-July-2023	26-July-2023

Abstract

The Covid-19 pandemic affects millions of people throughout the world. Everyone's life gets disturbed because of this pandemic. The Covid-19 vaccines are now freely available, but people still get feared because of the rumors spread by social media. Despite recommendations from experts, people show conceptions and perceptions regarding vaccines on social media platforms. The main objective of this study is to introduce the methodology to analyze the public's views regarding covid-19 vaccines using a publicly available worldwide Twitter dataset. In this study, we have used Natural Language Processing to analyze the sentiments. For sentiment analysis, we use Keras embeddings (deep Neural network) and Top2vec for topic modeling. This research will aid the government so that they can distinguish the major issues and provide preventive measures by taking the help of social media.

Keywords: covid-19, vaccine hesitancy, topic modeling, Twitter

1. Introduction

Social media platforms are considered a wealthy source of information for evaluating people's thinking and behaviors during any crisis. When the lockdowns and limitations are forced due to the spread of COVID-19, social media platforms are considered the public's meeting place for the people to share their conclusions and encounters relating to the covid-19 and covid-19 vaccines. But the newness of COVID-19 has led to the wrong and clashing information [1].

* Corresponding Author: Neelofar Memona (memon.neelo12@gmail.com)

A lot of well-known doctors and health workers have researched the covid-19 vaccines around the world and considered the vaccine safe and good for humans; said vaccines prevent people from various diseases, but still, there's more contention over the utilization of vaccines than other preventive measures (e.g., hand cleanliness, social distancing, sanitizing). These concerns incorporate fear of side impacts, instability around vaccine adequacy, and doubt about the science and the government [2].

Badly, social media platforms play a vital role in growing anti-vaccination developments and expanding vaccine hesitancy. People get feared by fake rumors spread by social media and aren't ready to get vaccinated, which is one of the biggest issues for the government and other health workers. Also, the public is not ready for the booster dose as well.

We shall investigate public perception regarding Covid-19 and shall seek to answer at least the following questions:

1. What are the emotions of the people concerning Covid-19 vaccines?
2. What are the main themes discussed on social media platforms regarding Covid-19 vaccination?

In this work, we will apply natural language processing to data found on the social media platform that is Twitter, where people share their covid-19 vaccine stories, and apply respective machine learning methods to datasets to observe the thinking of the public regarding Covid-19 vaccines. With the use of natural language processing (NLP), we will observe users' sentiment analysis with the help of their posts to be extracted from social media.

2. RELATED WORK

In the past, much research has been done to observe the public's sentiments toward Covid-19 vaccines. This section briefly describes the previous work for sentiment analysis and topic modeling.

Over the years, Natural language processing has been used for various applications: email filtering [3], language translation [4], text analytics [5], and more. Twitter data has also been used for various purposes: brand monitoring, sentiment analysis, competitor tracking, and more. More than 70000 Twitter data has been used for customer feedback for over a year. According to the ref, much Twitter data has been used for sentiment analysis. [6], have used Twitter data and find out six major themes that clearly show the public's sentiments and vaccine hesitancy. The major discoveries of this research are to bring detailed methodologies to move forward with Covid-19 vaccine ingestion, point out the key forms that require consideration within the arrangement of Covid-19 vaccination, and provide output on the hurdles and facilitators in current vaccination drives and permit for assist approach changes. The key discoveries outline three key parts of social media: observation and checking, a communication stage, and an assessment of government responses. This study was done when there was the first phase of covid-19; that's why it is not that accurate because the person's views cannot be the same over the year, but they are changing over time; there is variation in covid-19 vaccines as well, day by day new vaccines are introduced to overcome the pandemic.

The study ref. [7], also uses Twitter's data to find vaccine hesitancy among the people of the USA. Millions of tweets were analyzed using specific keywords like covid-19 and covid-19 vaccines. The study was done during the first phase of covid-19. Generally, the spread of data and people's views on social media platforms amid the

early organizing of the episode has significantly influenced individuals' convictions and state of mind towards covid-19 vaccines and their covid-19 vaccine choices. The research ref. [8], investigates transient advancement of distinctive feelings categories: Hesitation, covid-19 rollout, deception, and well-being impacts. They gather Twitter information from five nations: the UK, Brazil, the USA, India, and Australia. For vaccine hesitancy, all have used the same method (sentiment analysis and topic modeling).

A lot of research was also done on other social media platforms like Reddit, Facebook, etc. In ref. [9] they used Reddit data to find vaccine hesitancy. The polarity analysis of this study observed that fifty-six percent of the posts measured positive. In comparison, fifteen percent of the posts were neutral, and for the rest of the data, 28 percent of posts were negative. Machine learning models were also used to find vaccine hesitancy. The data fetched from the Reddit Platform is unsuitable for studying specific geographical areas. There may be some bots in collected samples of the data. In study ref. [1] they used logistical challenges (machine learning approach to find the vaccine hesitancy in the USA. They use the word2vec model to create dictionaries of different themes regarding different vaccinations for covid-19. They use news media to collect data and then create dictionaries.

The conclusion of this study shows that tweets communicating reluctance towards vaccines contain the most elevated notices of health-related impacts in all nations. They also demonstrate that the designs of aversion were variable over geographies and can assist in focusing on mediations. They watched a noteworthy alter within the direct patterns of categories like circumstance and satisfaction before and after the endorsement of vaccines. In ref. [10], they also used Reddit data to find vaccine hesitancy, but their research is limited to only a few cities in Canada; the thing is, Reddit is a public platform where anyone can comment all around the world, so it is tough to distinguish between the locations of the user.

In the Literature Survey, there is not only the involvement of social media, but also some researchers have collected data from their surrounding people using a questionnaire. In ref. [11], they collected data from nursing students using a questionnaire about vaccine hesitancy. When the data was collected, colleges were giving online instruction, and there were online classes, so the students did not completely understand the vaccine's significance can be considered a rule of impediment. Curfews were widespread, and students were candidly worn out as they went through as well long at home. It may trigger negative sentiments toward the vaccine and can be considered a limitation. In addition, low vaccination proficiency in students, the impact of social media, common uneasiness, and stretch due to Covid-19 may have influenced the considerations of students, and all these parameters may be a rule of restriction. In this consideration, the utilization of correlational shows and online collection of the information were among the other limitations.

We know that several researchers have analyzed covid-19 vaccine hesitancy. Sentiment Analysis, Topic Modelling, and Lexicon-based methods have been used for data understanding and analysis. Also, they took the data for a very short period. The second thing we have noticed is the research was done during the first phase of covid-19; the public's concerns might have changed over time as the development of vaccines progressed.

To the best of our knowledge, all the previous research had used Vader for sentiment analysis and LDA for topic modeling, which is not that accurate, so our major contribution is to find the perception of the public regarding covid-19 vaccine where people are hesitant or motivated to inoculate themselves.

3. Methodology

First, this section presents the data collection procedure and general statistics of the dataset. Then, it explains the computational techniques used to derive various data attributes. Then we explain how we clean the text using different techniques and then explain the methods for sentiment analysis and topic modeling. In the end, we explain the classification of data.

3.1 Data Collection

In this research, the publicly available Covid-19 Twitter dataset was used. There are almost 22455 tweets in the dataset. The dataset has various locations all over the world. The data contained different attributes, such as the name of the user, the location of the user, and the text. Some Examples from the dataset are given in Table 1.

Table 1: Some tweets from our dataset

Example of Tweets
Cannot see Corona anywhere? Everything is upside down, and in reverse, the vaccine creates the illness instead of the cure¦ https://t.co/HalpKnDRLj
l heard some lady on the radio say that the covid vaccine was like the flu vaccine where if you took it, you get benefited ¦ https://t.co/UqboNKvXFa
Get vaccinated corona vaccine prominent anti-vaxxer apologizing and telling people to avoid the vaccine.

3.2 Data Processing

In data science, pre-processing is one of the important steps. After collecting data, these techniques were used for data cleaning, as shown in Figure 1, to get the required results [12]. Pre-processing techniques include the Conversion of tweets from uppercase to lowercase and removing punctuation, URLs, and stop-words from the tweets. In the end, the Lemmatization and stemming of the tweets were done.

3.3 Sentiment Analysis

It is an approach in Natural Language Processing that identifies the sentiment of the data [13]. In this study, we have used the Keras embedding model for text classification and sentiment analysis. We have used two approaches of the word embedding model to compare the accuracy of the data. The first approach is flattened word embedding, and the second is word embedding averaged. At first, we load the data and vectorize it in the train and test dataset, then tokenize it, as shown in Figure 2.

Tweets before removal of hashtags, usernames & hyperlinks

```
                                                  text  \
0      Great thing about the #German language are com...
1      @mcezeh1 @NCDCgov Hope you're wearing your mas...
2      Welcome to "THE SHIT SHOW" that is Donald Trum...
3      I've never seen Dr. Fauci so happy! He can tal...
4      #GetVaccinated #CoronaVaccine \n\nA Prominent ...
```

Tweets after removal of hashtags, username & hyperlinks

```
                                                cleanText
0      great thing german language compound words new...
1      mcezeh ncdcgov hope wearing mask washing hands...
2      welcome shit show donald trump invited amp buy...
3      never seen fauci happy talk freely science bac...
4      getvaccinated coronavaccine prominent anti vax...
```

Figure 1: Removal of Hashtags, Usernames, and Hyperlinks

Feat Vector for text

Text: Israel trades Pfizer vaccine doses medical data covid coronavac corona vaccine

```
Feat Vec for Text: [ 149 1780   20    2   23  106  168    3  391   25    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0]
```

Figure 2: Vectorization

In the first approach, we used three layers. A single embedding layer and two dense layers train the model to check the validation. Using these layers, we classified data; the classifiers divided the data into three categories Negative, Neutral, and positive, as shown in Table 2.

Table 2: Experimental Results

Classification Report of Approach 1				
Predictions	Precision	recall	f1-score	support
Negative	*0.73*	*0.45*	*0.56*	*553*
Neutral	*0.82*	*0.94*	*0.88*	*1972*
Positive	*0.91*	*0.86*	*0.88*	*1721*

We used the LIME algorithm using the Python lime library for sentiment analysis. LIME stands for Local Interpretable Model-agnostic Explanations. This library trains the models to explain the predictions of every individual comment. This library can be applied to any Machine Learning model. In our study using LIME, we compare the predicted and actual text to analyze whether the text is positive, negative, or neutral, as shown in Figure 3. In the following example, the given text is neutral, and the prediction of LIME is also unbiased.

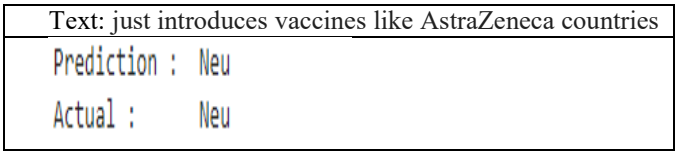


Figure 3: LIME Algorithm Example

Coming to the second approach is word embedding averaged. Like the previous network, three layers are used (one embedding layer and two dense layers), and the whole process is the same; only the difference is that the output of the embedding layer is averaged at the token index.

Table 3: Classification of Data

Classification Report of Approach 2				
Predictions	Precision	recall	f1-score	support
Negative	*0.85*	*0.65*	*0.74*	*553*
Neutral	*0.86*	*0.96*	*0.91*	*1972*
Positive	*0.92*	*0.88*	*0.90*	*1721*

Now we compare the accuracy of both approaches as shown in Figure 4.

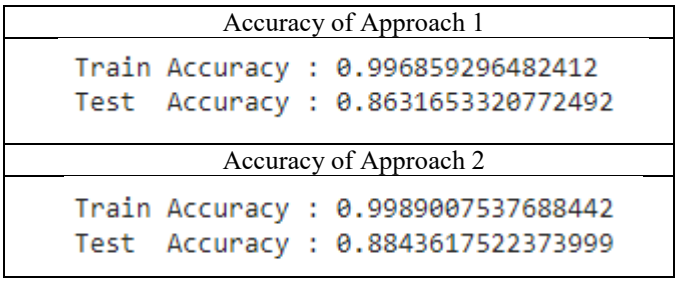


Figure 4: Accuracy comparison of both approaches

From the accuracy, we can notice that 2nd approach is slightly better than the first one. We also create a confusion matrix in Figures 5 and 6 to check the performance.

3.4 Topic Modelling

It is a machine-learning approach for discovering topics in a document. It helps in discovering hidden topical patterns that are present across the document. There are a lot of algorithms to apply topic modeling. In this study, we have used the Top2vec model for topic extraction. The Top2Vec algorithm is considered one of the straightforward ways to perform topic modeling. It is not only limited to topic modeling but is also used for semantic relation searches in a document. This algorithm automatically recognizes a text document's topics, generating jointly embedded topics and word vectors. The most important thing about this algorithm is atomic

features, and it also has a lot of functions that can work on both short and long text. So, initially, we installed the Top2Vec library from Python and uploaded the pre-processed data. It searches out 24 topics from the dataset. Then we create joint embedding of the document and word vectors. Once done, the algorithm finds dense clusters of documents to identify which word attracted the document. In the end, we create the word clouds of the topics, as shown in Figure 8.

4. Results

We first took the publicly available data and performed the pre-processing techniques. After that, we used the Keras word embedding model to generate the sentiments of the tweets, namely positive, negative, and neutral. We compared the accuracy of the two approaches flattened and averaged. Then we further analyze the Top2vec model to generate the topics of positive and negative tweets and examine the hot topics discussed in the tweets.

4.1 Confusion Matrix for test prediction of approaches 1 and 2

We have used two approaches and checked both accuracies; we also draw a confusion matrix to elaborate the results further.

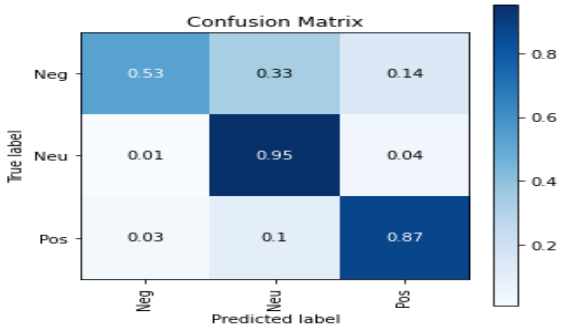


Figure 5: Confusion Matrix for Approach 1

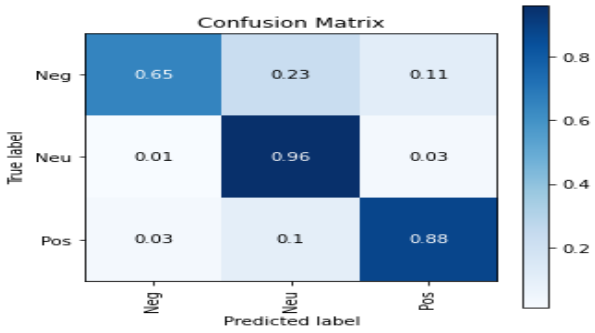


Figure 6: Confusion Matrix for Approach 2

4.2 LIME Algorithm for Sentiment Analysis

In the Keras word embedding model, we have used the LIME algorithm for the prediction of the sentiments of the tweets. This algorithm compares the actual and predicted sentiment of the individual tweet and shows how much percentage of the words is negative, positive, or neutral. It also highlights the words which indicate the sentiments, as shown in Figure 7. This is one of the best sentiment analysis techniques among others.

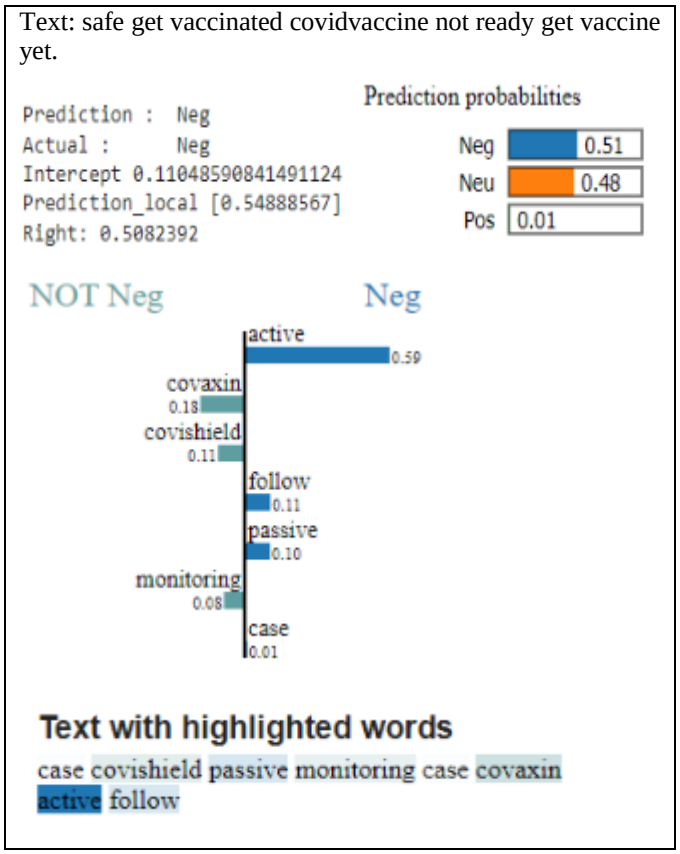


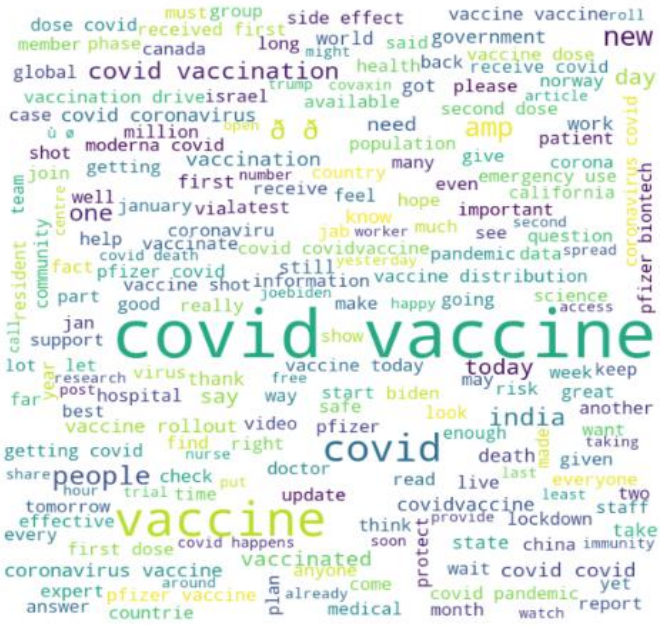
Figure 7: LIME Algorithm

4.3 Topic Modelling

We used the Top2vec algorithm to extract topics from the data and generate the word clouds.

Acknowledgments

This research was done at the Department of Telecommunication Engineering MUET Jamshoro. All the authors have equal contributions to it.



200

Figure 8: Wordcloud

References

[1] Shantanu Dutta, Ashok Kumar, Moumita Dutta, and Caolan Walsh. Tracking covid-19 vaccine hesitancy and logistical challenges: A machine learning approach, 6 2021

205 [2] Nobutoshi Nawa, Shigetoyo Kogaki, Kunihiko Takahashi, Hidekazu Ishida, Hiroki Baden, Shinichi Katsuragi, Jun Narita, Keiko Tanaka-Taya, and Keiichi Ozono. Analysis of public concerns about influenza vaccinations by mining a massive online question dataset in Japan. Vaccine, 34:3207–3213, 6 2016.

[3] X. Carreras and L. Màrquez, "Boosting trees for anti-spam email filtering," 2001, arXiv: cs/0109015.
210 [Online]. Available: https://arxiv.org/ abs/cs/0109015

[4] M. Abbaszade, V. Salari, S. S. Mousavi, M. Zomorodi, and X. Zhou, "Application of Quantum Natural Language Processing for Language Translation," in IEEE Access, vol. 9, pp. 130434-130448, 2021, DOI: 10.1109/ACCESS.2021.3108768

[5] R. Valdez-Almada, O. M. Rodriguez-Elias, C. E. Rose-Gomez, M. D. J. Velazquez-Mendoza, and S.
215 Gonzalez-Lopez, "Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions: Generating Knowledge Profiles from Resumes," 2017 5th International Conference in Software Engineering Research and Innovation (CONISOFT), 2017, pp. 97-106, DOI: 10.1109/CONISOFT.2017.00019

[6] Chandrasekaran R, Desai R, Shah H, Kumar V, Moustakas Examining Public Sentiments and Attitudes
220 Toward COVID-19 Vaccination: Infoveillance Study Using Twitter PostsJMIR Infodemiology 2022;2(1): e33909URL:https://infodemiology.jmir.org/2022/1

[7] Li Crystal Jiang, Tsz Hang Chu, and Mengru Sun. Characterization of vaccine tweets during the early stage of the covid-19 outbreak in the united states: A topic modeling analysis. JMIR Infodemiology, 1: e25636, 9 2021

[8] Harshita Chopra, Aniket Vashishtha, Ridam Pal, Ananya Tyagi, and Tavpritesh Sethi. Mining Trends of COVID-19 Vaccine Beliefs on Twitter with Lexical Embeddings

[9] Chad A. Melton, Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. Public sentiment analysis and topic modeling regarding covid-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14:1505–1512, 10 2021.

[10] Janessa Griffith, Husayn Marani, and Helen Monkman. Covid-19 vaccine hesitancy in Canada: Content analysis of tweets using the theoretical domains framework. Journal of Medical Internet Research, 23, 4 2021.

[11] Akgün Yeşiltepe, Sinan Aslan & Semra Bulbuloglu (2021) Investigation of perceived fear of COVID-19 and vaccine hesitancy in nursing students, Human Vaccines & Immunotherapeutics, 17:12, 5030-5037, DOI: 10.1080/21645515.2021.2000817

[12] Bhaya, Wesam. (2017). Review of Data Pre-processing Techniques in Data Mining. Journal of Engineering and Applied Sciences. 12. 4102-4107. 10.3923/jeasci.2017.4102.4107.

[13] M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials, and Electronic Engineering (IC4ME2), 2019, pp. 1-4, DOI: 10.1109/IC4ME247184.2019.9036670.