International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentiment Analysis Using Machine Learning Techniques

Garlapati K V S Sai Komal[1], *Dr. Sagar Dhanraj Pande*[2]

[1]*Student , School of Computer Science and Engineering, Vit-ap University, Guntur, Andhra Pradesh*
[2]*Faculty, School of Computer Science and Engineering, Vit-ap University, Guntur, Andhra Pradesh*

ABSTRACT

The aim of this research paper is to analyze the tweets which are in response to the COVID-19 vaccine Pfizer-BioNtech. There are other vaccines for this virus like Moderna, and AstraZeneca, but we choose the tweets of the Pfizer vaccine. The data which we use to extract the sentiment are based on Twitter tweets collected using the tweepy package in Python. Tweets contain information regarding people's opinions about the vaccine.

Keywords: Support Vector Machine (SVM), Naïve Bayes, Random Forest, Logistic Regression, TF-IDF vectorizer

1. Introduction

1.1 Effect of COVID-19

Pfizer is a vaccine for the coronavirus which first originated in Wuhan, Hubei Province, the People's Republic of China, in December 2019. This virus spread across the globe rapidly. People were to be followed by curfew and stuck in their places. The government has taken measures to keep the public safe, such as quarantines, lockdowns, and staying home. This eventually made it difficult to conduct surveys or to gather people's opinions[1]. Social media is the only access available to contact the outside world. Some researchers began to use social media to analyze public sentiments because the information can be acquired more efficiently from social media at a lower cost[2]. Twitter is one of the largest social networking platforms which is a good data source. It is the major source people used to express their opinions and emotions. It is a platform where the public expresses their reviews openly, hence is a valuable source for research in sentiment analysis. This project analysis the tweets got on the Pfizer vaccine

1.2 Sentiment Analysis Process

Sentiment Analysis is the study of opinions and emotions in a particular context. There are two technical terms used SA(Sentiment Analysis) and OM(Opinion Mining) which can be used interchangeably but [3] some say that both differ. SA is about emotion and OM is related to opinions. SA is done on a sentence or a paragraph to extract its sentiment. There are stages in sentiment analysis which are discussed below.

The raw data that is the original tweets are preprocessed using NLP. This processed data is used to determine its polarity and subjectivity. Based on these values we use various machine learning techniques for classification. Sentiment in the data is classified as positive, negative, and neutral. We used logistic regression to classify the data into the above classes.

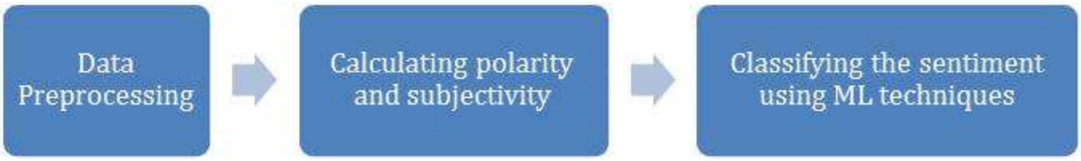

Data Preprocessing
Calculating polarity and subjectivity
Classifying the sentiment using ML techniques

Fig 1.1 Flow chart of Analysis

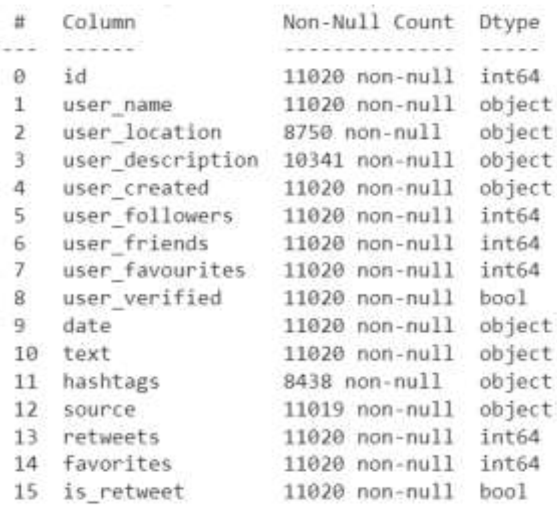
2. LITERATURE REVIEW

A study in [4] states that tweeting about vaccination-related health information with the help of influential social media users could reach a wider audience. According to [5], posts on the internet reveal the public's interest and attitude towards vaccines. They claimed that the public are more interested to engage with negative responses. In a study that looked at tweets on the flu vaccine in 2018, Sarker and Gonzalez [6] discovered that sentiment analysis can be a useful method for identifying issues and concerns about vaccine acceptability. Additionally, other research has explicitly examined tweets about the Pfizer vaccination. Sentiment analysis was employed in a study by Ahmed et al.[7] in 2021 to examine tweets on the Pfizer vaccine, and they discovered that the majority of them were supportive and positive. The investigation did discover a sizable amount of tweets that were critical of vaccines and raised security issues. Similar to this, a study conducted in 2021 by Nabais et al [8]. employed sentiment analysis to examine tweets on the Pfizer vaccine and discovered that while most of them were supportive, there were also some critical ones about the vaccine's effectiveness and distribution. The study in [9] found that throughout time, people's opinions and feelings concerning COVID-19 changed from being positive (at the start of the epidemic) to negative (by the end of 2021).

3. METHODOLOGY

3.1 Data Preprocessing

The dataset consists of many unwanted features which not used in the analysis. We can remove unwanted features. There are 16 features, the necessary one is only the text column which consists of the raw tweets. There is much-unwanted information in the text column (ex. Hashtags, mentions, emojis) which needs to be cleaned. Initial columns are



```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                11020 non-null  int64
 1   user_name         11020 non-null  object
 2   user_location     8750 non-null   object
 3   user_description  10341 non-null  object
 4   user_created      11020 non-null  object
 5   user_followers    11020 non-null  int64
 6   user_friends      11020 non-null  int64
 7   user_favourites   11020 non-null  int64
 8   user_verified     11020 non-null  bool
 9   date              11020 non-null  object
 10  text              11020 non-null  object
 11  hashtags          8438 non-null   object
 12  source            11019 non-null  object
 13  retweets          11020 non-null  int64
 14  favorites         11020 non-null  int64
 15  is_retweet        11020 non-null  bool
```

Table.1 About Dataset

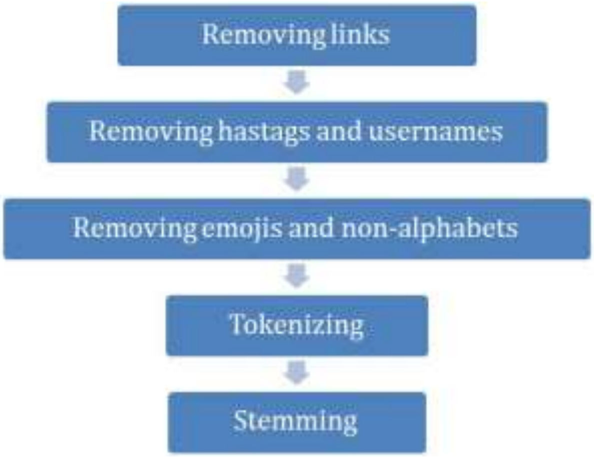
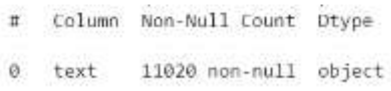

Removing links
Removing hastags and usernames
Removing emojis and non-alphabets
Tokenizing
Stemming

Fig 3.2 Flow chart of Preprossing

The only required column is the ‘text’ column. Dropping the unwanted columns and after checking for null values, no null values are found.



```
 #   Column  Non-Null Count  Dtype

 0   text    11020 non-null  object
```

Fig 3.3 Number of columns

By using the regex module from Python, hashtags and usernames, and other data are removed. Tokenization in NLP is a process of splitting sentences and paragraphs into smaller units, such that they are more meaningful. Applying tokenization to the tweets, and stemming the words. Stemming is reducing the words to their base form by removing the suffixes and prefixes.

3.2 Polarity and Subjectivity

Polarity is a value that returns a numerical score ranging from -1 to 1 based on the content of a sentence. It classifies the emotion of the text into positive(+1), negative(-1), or neutral(0). Subjectivity is also a value ranging from 0 to 1 where 0 is related to more objectivity and 1 refers to more subjectivity. Objectivity is more related to facts while subjectivity is more related to opinions and feelings. Text Blob can be used to calculate these both. It uses a lexicon-based approach to calculate the polarity.

3.3 Classifying the Sentiment

The text, polarity, and subjectivity columns have been added to the dataset. Based on the polarity value, the sentiment column is added to the dataset. Neutral if the polarity is equal to zero. If the positive emotion is more than zero and the negative sentiment is less than zero. The classifiers are trained using this labeled data. The classifiers listed below are based on logistic regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM). Using train_test_split from sci-kit-learn, the dataset is divided into training and testing datasets for each model. TF-IDF vectorizer to translate the text into numerical numbers

3.3.1 Support Vector Machine

It is a powerful supervised learning model. We will use this to classify these three classes. Since there are three classes we will use the multiclass SVM model.

3.3.2 Naïve Bayes

Naïve Bayes is a supervised machine learning model which uses probabilistic classifiers. Probabilities are calculated using the target variable. Since there are three classes in the target variable we can use a Multinomial classifier for this dataset.

3.3.3 Random Forest

It is also a supervised model based on ensemble learning. It has multiple decision trees. More the number of trees greater the accuracy.

3.3.4 Logistic Regression

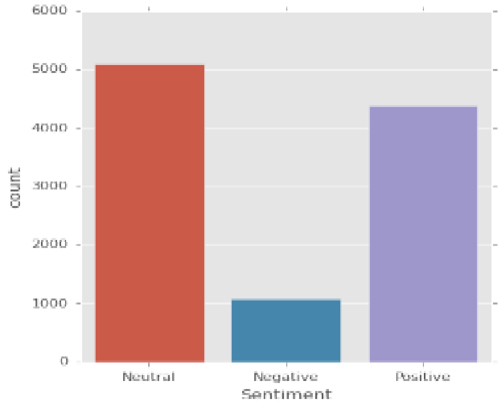
By the name, it is not a regression algorithm. It classifies using an ‘s’ shaped curve. It uses the Linear regression values and classifies them as 0 or 1.

4. RESULTS AND DISCUSSIONS

The accuracy of each model is listed below. Comparatively, SVM has the highest accuracy of all.

Algorithm	SVM	Naïve Bayes	Random Forest	Logistic Regression
Accuracy	80.7582	72.8436	79.3364	77.9146

Visualizing the total tweets in the three classes, there are very less negative tweets compared to the positive and negative tweets. Fig 4.1 and Fig 4.2 represents the bar graph and pie chart of the total number of tweets in three classes.



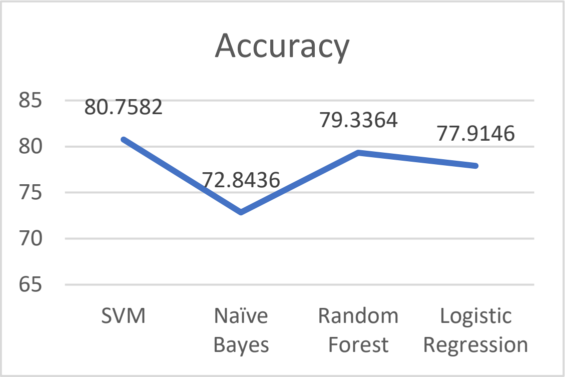


Fig 4.1 Bargraph of tweets

Fig 4.2 Line plot of accuracies

Word cloud of the three classes are listed in Fig 4.3, Fig 4.4, and Fig 4.5

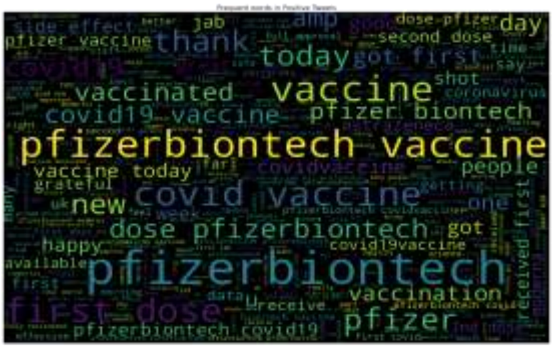
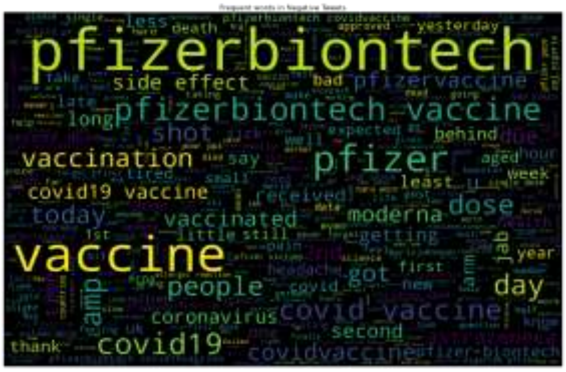


Fig 4.3: Frequent words in positive tweets



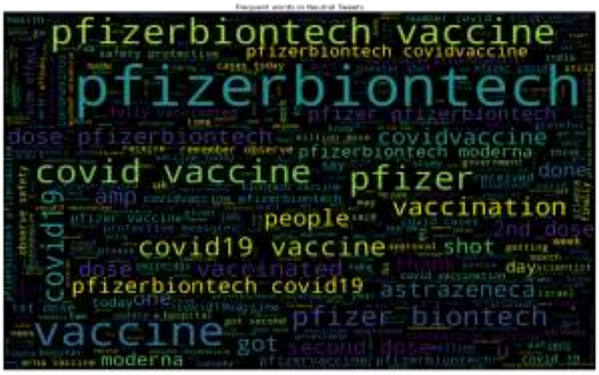


Fig 4.4: Frequent words in negative tweets

Fig 4.5: Frequent words in neutral tweets

5. CONCLUSION

In this paper, we have successfully implemented various machine learning techniques on the Pfizer vaccine dataset, classifying them into positive, negative, and neutral tweets. SVM model has the highest accuracy of 80% The pandemic has taken many lives of the public. With the arrival of vaccines which is not an actual cure but can prevent virus attack. This has an immediate reaction in public and is expressed through social media. What was this reaction actually, is it happy or angry? This is the sentiment and we classified that the majority of people are satisfied and very few are unsatisfied.

REFERENCES

[1] Insight from NLP Analysis: COVID-19 Vaccines Sentiments on Social Media Tao Na∗ , Wei Cheng∗ , Dongming Li∗ , Wanyu Lu∗ , Hongjiang Li∗ ∗Department of Computer Science, University of Manchester, Manchester, U

[2] Hussain, A.; Tahir, A.; Hussain, Z.; Sheikh, Z.; Gogate, M.; Dashtipour, K.; Ali, A.; Sheikh, A. Artificial intelligence–enabled analysis of public attitudes on Facebook and twitter toward COVID-19 vaccines in the united kingdom and the united states: Observational study. J. Med. Internet Res. 2021, 23, e26627.

[3] Survey on mining subjective data on the web | SpringerLink

[4] A study of two vaccine-related twitter datasets. Perm. J. 2018, 22 , 17-138. Blankenship, E.B.; Goff, M.E.; Yin, J.; Tse, Z.T.H.; Fu, K.W.; Liang, H.; Saroha, N.; Fung, I.C.H. Sentiment, contents, and retweets:

[5] Facebook and Twitter vaccine sentiment in response to measles outbreaks. Health Inform. J. 2019, 25, 1116–1132. Deiner, M.S.; Fathy, C.; Kim, J.; Niemeyer, K.; Ramirez, D.; Ackley, S.F.; Liu, F.; Lietman, T.M.; Porco, T.C.

[6] Portable automatic text classification for adverse drug reaction detection via multi-corpus training Abeed Sarker∗, Graciela Gonzalez∗

[7] Ahmed F, Nabais MF, Zafar A, Soomro TR, Merchant AA, Al-Garadi MA. Sentiment analysis and topic modeling of tweets related to Pfizer-BioNTech COVID-19 vaccine: a cross-sectional study on multiple machine learning techniques.

[8] Nabais MF, Ahmed F, Soomro TR, Zafar A, Merchant AA, Al-Garadi MA. Understanding the public's sentiment towards the Pfizer-BioNTech COVID-19 vaccine: a cross-sectional study based on Twitter data analysis. Journal of medical Internet research.

[9] Alhuzali H, Zhang T, Ananiadou SEmotions and Topics Expressed on Twitter During the COVID-19 Pandemic in the United Kingdom: Comparative Geolocation and Text Mining Analysis J Med Internet Res 2022;24(10):e40323 URL: https://www.jmir.org/2022/10/e40323 DOI: 10.2196/40323