American Journal of Electronics & Communication

Srestha Sadhu et. al., American Journal of Electronics & Communication, Vol. III (2), 21-29

Sentiment Analysis on COVID-19 Twitter Data

Srestha Sadhu[1]
Department of Computer Science and Engineering
University of Engineering and Management,
Kolkata.
srestha.sadhu@gmail.com

Varsha Poddar[2]
Department of Computer Science and Engineering
University of Engineering and Management,
Kolkata.
varsha.poddar@gmail.com

Puja Paul[3]
Department of Computer Science and Technology
University of Engineering and Management,
Kolkata.
puja.paul.uemk.cst.2023@gmail.com

Sheraly Hansda[4]
Department of Computer Science and Technology
University of Engineering and Management,
Kolkata.
sheraly.hansda.uemk.cs.2023@gmail.com

Rimpa Saha[5]
Department of Computer Science and Technology
University of Engineering and Management,
Kolkata.
rimpa.saha.uemk.cs2023@gmail.com

Angira Chakraborty[6]
Department of Computer Science and Engineering
University of Engineering and Management,
Kolkata.
angira.chakraborty.uemk.cse.2023@gmail.com

Titiksha Paul[7]
Department of Computer Science and Engineering
University of Engineering and Management,
Kolkata.
titiksha.paul.uemk.cs.2023@gmail.com

Anushree Mondal[8]
Department of Computer Science and Information Technology
University of Engineering and Management, Kolkata.
anushree.mondal.uemk.csit.2023@gmail.com

Abstract

Coronavirus first appeared in December 2019 in Wuhan, China which eventually lead to a catastrophic impact all over the world. The entire world had been fighting this pandemic and expressing their feelings, sharing their opinions on various social media platforms. Substantially Twitter had been the better medium to express their opinion and share updates about the situation. This paper analyzes the sentiments of the public based on positive, negative, and neutral tweets. This analysis eventually helps in the prediction of the covid-19 situation in the world. The dataset was collected from Kaggle which was uploaded by Gabriel Preda containing more than 1,70,000 tweets. Based on these tweets posted on Twitter a sentiment analysis was performed. Data Collecting, Data cleaning (Removing URL, #, @ and various types of punctuation), Tokenization, Stemming, Removing stop- words were performed, and to find the polarity, two types of analyzers were used that is TextBlob and Afinn. 1,79,108 tweets were manually analyzed and comparing it with both the analyzers shows Afinn is more accurate. To evaluate the accuracy a few Machine learning Algorithms had been applied (Logistic Regression, Naive Bayes, Decision Tree, and Linear Regression) for predicting the sentiment of the tweet.

Keywords: COVID-19, Sentiment analyzers, Logistic Regression, Decision Tree, Naïve Bayes, Linear Regression.

1. Introduction

Social media platforms such as Facebook, Twitter, and YouTube, provide us with information known as social data. This data is used for analyzing and predicting the future of various fields. Twitter had been chosen for this research as the platform to convey the thoughts and emotions of people worldwide. Twitter, a social networking and an online news site is used to communicate in short messages called tweets. Coronavirus known as COVID-19 is one of the most recently discussed topics in the world which has impacted negatively the day-to-day life of people. Many have lost their dear ones, lost their jobs, traveling is still restricted at some parts of the world and businesses are running at losses. Hence people chose to share their feelings in the form of help, gratitude, positively and expressed their pain with the world on social media platforms since the beginning of the pandemic, March 2020. Followed by the current situation this work focuses on the sentiments of people by collecting tweets from an already available dataset taken from Kaggle uploaded by Gabriel Preda. This dataset consists of data, dating from 25.7.2020 to 29.8.2020. Based on the positive, negative, and neutral tweets the impact of the coronavirus on the minds of people had been analyzed.

Sentiment Analysis on "COVID-19" Twitter data had been performed by using various Analyzers and Machine Learning Algorithms to find the polarity as well as the accuracy of the following Twitter dataset. Python had been chosen since it provides numerous libraries to access social media platforms like Twitter.

The dataset has been collected and cleaned through data cleaning, tokenization, stemming, and removing stop-words. For performing sentiment analysis both TextBlob and Afinn analyzer had been used to evaluate the polarity and then compared both the results. It has been found that using TextBlob the accuracy is more compared to Afinn but after evaluating the sentiments of each tweet manually it was observed that Afinn is more precise. To find the accuracy, various machine learning algorithms such as Logistic Regression, Naïve Bayes, Decision Tree, and Linear Regression were used. The results showed more neutral and positive tweets than negative tweets. Based on this, the maximum accuracy of 96.65% using the Logistic Regression algorithm was obtained whereas the other algorithms gave 82.09%, 95.51%, and 70.99% respectively.

The main objective is to analyze the sentiment of people due to covid-19 between July, 2020- August, 2020. During this period, coronavirus was in its initial phase and the Centers for Disease Control and Prevention (CDC) were ongoing research for the vaccine to develop and manufacture them. With rapid research development, people were expecting that the world would come up with a solution for the situations to get better globally. Hence, people were spreading awareness worldwide and holding campaigns for free RT-PCR tests. Therefore, it is expected to get more neutral and positive tweets than negative tweets.

2. Related Works

Many researchers have performed elaborate studies to show the effectiveness of Machine learning in sentiment analysis of people throughout the world fighting with the Virus. COVID- 19 has caused a global crisis which has changed the perception of the world and compelled people to deal with the large scale disaster caused by it which has also impacted people, psychologically. Data from Twitter contributes and helps to discover sentiments of the people in various cases when the world is facing a pandemic situation .

This section of the paper covers several important papers related to COVID-19 sentiment analysis dated from 2020 to 2021 which were used as references. Researchers of [1] [4] [5] [6] [7] [8] [11] have performed Public Sentiment Analysis on Twitter Data since the outbreak of COVID-19 in 2020 where preprocessing is done using NLTK library and TextBlob analyzer is used to analyze the polarity and subjectivity of tweets.

A survey conducted by Abdul Aziz et al.[2] in 2021, indicated the use of Naïve Bayes Classifier (NBC) and Logistic Regression (LR) classification methods of machine learning. In 2020, Stanislaw Wrycza [3] also published a research work using Naïve Bayes. In 2020, Drias et al. [9] performed a study using the lexicon-based approach. In 2021, Gupta et al.[10] published research work to classify the data accurately. Eight different classifiers are used (Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, LinearSVC, AdaBoost Classifier, Ridge Classifier, Passive Aggressive Classifier, and Perceptron).

In 2021, Supriya Raheja[12] published a research work that visualizes textual data through WordCloud and categorizes Tweets into notions like positive, negative, and neutral. In 2021, Piyush Ghasiya [13] published a study by creating a labeled dataset using unsupervised machine learning methods. In 2020, Machuca et al. [14] performed research work aiming to deduce whether the sentiment is positive or negative by applying machine learning algorithms and NLP techniques. In 2021, Amir Hussain [15] worked on a research paper where social media data is largely unstructured, but natural language processing (NLP) and machine learning (ML) is used. In 2021, Carol Shofiya [16] used the Hybrid Approach by employing the sentiment polarity and then applying the SVM algorithm, for classification and analysis. In 2021, Amir Rehman [17] disseminated a research work using CNN and D.N.N. which were the most suitable classification techniques to detect sentiment, followed by SVM, Random Forest, K-NN, and L.S.T.M.

By going through all the papers it is found that TextBlob analyzer and Naïve Bayes algorithm are used the most. By gathering information from all these papers and keeping them as references, we are inspired to proceed further with our study. The organization of the paper is as follows: Section 3 describes the subject dataset while Section 4 describes elaborately the implementation mechanism under which the data preprocessing and application model is explained. In Section 5, the Results and Findings of the research are given elaborately. Section 6 and Section 7 give the conclusion and future scope of the paper, respectively.

3. Subject Dataset

In this research, one subject dataset with 13 attributes has been chosen upon which all the necessary operations have been implemented. This dataset has data records from 25/7/2020 to 29/8/2020. In this period, 179108 tweets have been tweeted by the people regarding the COVID-19 pandemic. According to the graph of that time, corona cases were less than other times. Our work is to find out the sentiment of people over this period.

Name of the Attribute	Description of the Attribute
user_name	User identification
user_location	From where the user belongs
user_description	Explains what the user does
user_created	Date & time when user account was created
user_followers	Number of followers of user on twitter
user_friends	Number of user friends on twitter
user_favourites	Number of favourite users
user_verified	Verification of user account
date	Date & time of post
text	The comment
hashtags	A keyword explains a specific topic
source	From where the tweet was posted
is_retweet	Whether tweet has been reposted

Table 1 . List of all 13 attributes with their description

4. Implementation Mechanism

4.1 Data Preprocessing:

The prime focus of pre-processing is to clean the data and prepare for further implementations. Firstly all the tweets have been cleaned by removing contains like hashtags, user handles, white spaces, URLs, punctuations, and stop-words. Then tokenization and stemming have been done on the cleaned tweet.

Now two analyzers named as TextBlob and Afinn were applied to estimate the polarity scores as well as the sentiment of each tweet in the dataset. The cleaned tweets from the previous step were subjected to multiple evaluation models using TextBlob and AFINN.

TextBlob library supports complex analysis and operations on textual data. Here it is used to get polarity and subjectivity scores associated with each tweet.

AFINN is one of the most popular lexicon-based approaches used for sentiment analysis which is suitable for many languages. It contains a method called score() which takes a sentence as input and returns polarity score as output. It also contains more than 3300 words with a polarity score associated with each word. A generalized score of polarity was found for each tweet. Both analyzers return a value in the range [-1to+1] where +1 implies Extreme Positive Polarity,
-1 implies Extreme Negative Polarity and 0 implies neutral.

Finally, the dataset was ready to split into the form of training and testing data. Here 80% of the record was selected as train data while the rest 20% was considered as test data.

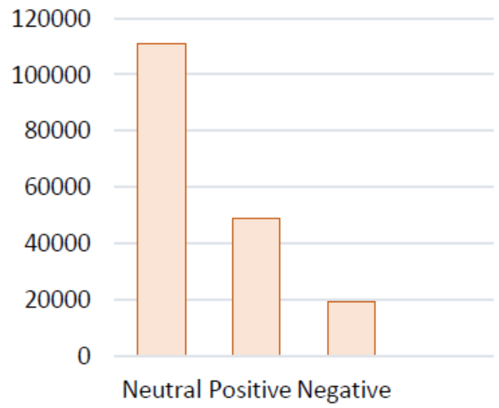

Fig 1. Using TextBlob Analyzer

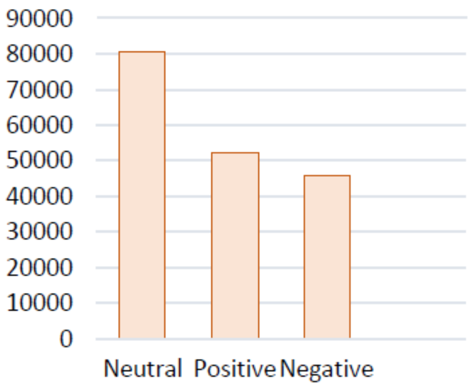

Fig 2. Using Afinn Analyzer

Fig 1 shows the diagrammatical representation of the number of neutral tweets that is 111110, positive tweets are 48904 and negative tweets are 19094.

Fig 2 shows the diagrammatical representation of the number of neutral tweets that is 80798, positive tweets are 52433, and negative tweets are 45877.

It has been found that using TextBlob analyzer the number of neutral tweets is more compared to Afinn but after evaluating the sentiments of each tweet manually it was observed that Afinn is more precise and accurate therefore TextBlob analyzer has been discarded for further application.

4.2 Application of Models:

In this research, for predicting the sentiment of the tweets various supervised machine learning algorithms were implemented:

Logistic Regression: Logistic Regression is a binary classification algorithm, used to predict a dependent variable based on a set of independent variables such that the dependent variable is categorical. The name is "regression" because its working technique is quite similar to linear regression.

Decision Tree: It is a supervised machine learning algorithm used to solve both classification and regression problems. It is a tree that helps us in decision-making purposes.

Naïve Bayes: It is a supervised machine learning algorithm based on the Bayes theorem used for solving classifications problems based on the probability of an object. It assumes all the features are conditionally independent but in the case of a real dataset no features are conditionally independent but they can be close.

Linear Regression: Linear Regression is a supervised machine learning approach that is used for predictive analysis. It makes predictions for a continuous or numeric variable.

For all the above cases accuracy values were obtained from the confusion matrix. **Table 2** tabulates all the accuracy values. For each run, the error rate was also determined and the time taken to complete the execution was also noted down.

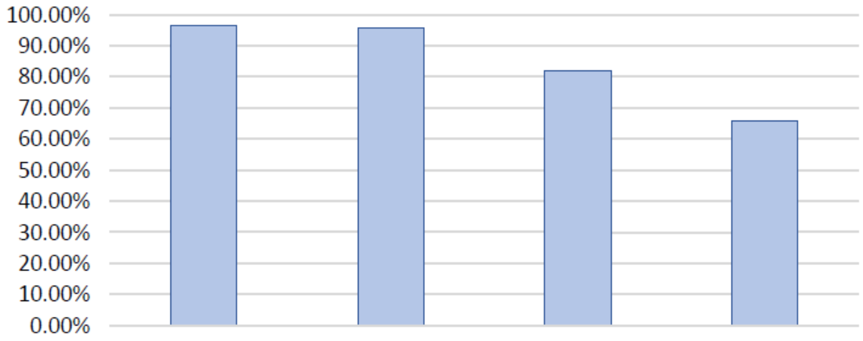

Fig 3. Accuracy of different models using AFINN Analyzer

ALGORITHM	**Logistic Regression**	**Decision Tree**	**Naïve Bayes**	**Linear Regression**
Accuracy	96.65%	95.51%	82.09%	70.99%
Error rate	0.04	0.05	0.18	0.3
Training Time	7.842s 17.417s 8.754s	59.445s 115.064s 71.474s	0.037s 0.097s 0.037s	19.574s 31.347s 19.66s
Prediction Time	0.0s 0.007s 0.03s	0.06s 0.063s 0.098s	0.004s 0.007s 0.072s	0.002s 0.002s 0.0s

Table 2. Using AFINN Analyzer

5. Results and Findings

A dataset consisting of 1,79,108 tweets was collected for this research. As shown in **Fig 1** and **Fig 2**, the number of neutral, positive, and negative tweets using TextBlob is 111110, 48904, and 19094, respectively, while the number of neutral, positive, and negative tweets using Afinn is 80798, 52433, and 45877.

The sentiment of the tweets was assessed in the first half using two separate analyzers – TextBlob and Afinn. It was discovered that the TextBlob analyzer produces more neutral tweets than Afinn, however after manually analyzing the sentiments of each tweet, it was discovered that Afinn is more precise and accurate, hence TextBlob analyzer has been eliminated for further use.

In the second phase, four algorithms were applied on the Afinn analyzer which is as follows: Logistic Regression produced the highest accuracy of 96.65%, using Decision Tree the accuracy was 95.51%, by Naïve Bayes the accuracy was 82.09%, and using Linear Regression accuracy was 70.99%.
For each run the error rate has been determined and also the training time and prediction time have been noted down as shown in **Table 2**.

6. Conclusion

The study of Sentiment Analysis on Twitter Data related to COVID-19 was presented in the research paper. The Prime focus of this research is to find out Positive vs Negative vs Neutral sentiment. It was observed that the highest sentiment from all the tweets, eventuated for Neutral. Almost all countries were expressing their feelings about COVID-19 on various social media platforms, but most of the tweets were obtained from Twitter Web App on #COVID19 for this paper. It had been concluded that Sentiment Analysis using Afinn gives the most accurate result. It had been compared with the manual analysis that Afinn is better than TextBlob. Moreover, Afinn contains 3300+ words with a polarity score associated with each word, and using this library package one can even find the sentiment score of different languages as well.

This work would beneficiate analyzing the sentiments of the people during the pandemic COVID-19 using different types of algorithms. After comparing the results of all the algorithms that are Logistic Regression, Decision Tree, Naive Bayes, and Linear Regression, it had been evaluated that the result of Logistic Regression is returning the best result with the highest accuracy. This is because the Logistic Regression algorithm has low variance means less error in test data so less chance of overfitting. This study provided a good analysis of sentiments and from this study, it can be said that the people's reactions vary day to day from posting their feelings on social media specifically on Twitter.

7. Future Scope

In the future, the aim is to collect the tweets using Twitter API and plan to create a dataset, as well as use other machine learning algorithms like Hybrid algorithm and then compare the result of the labeled dataset with the result of sentiment analysis that had been performed in this research.
This model can be further taken to new possibilities of Emotion analysis rather than having Positive, Negative, and Neutral tweets. This sentiment analysis model can also be applied to examine the shifting emotions and feelings of individuals and to check if there are noticeable changes over time in them.

References

1. Kausar, Mohammad Abu, Arockiasamy Soosaimanickam, and Mohammad Nasar. "Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak.” https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Public+Sentiment+Analys is+on+Twitter+Data+during+COVID19+outbreaks&btnG=&oq=Public+Sentiment+Ana lysis+on+Twitter+Data+during+COVID-19+Outbreak
2. Abdulaziz, Manal, et al. "Topic based Sentiment Analysis for COVID-19 Tweets." (2021), https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Topic+based+Sentiment+ Analysis+for+COVID-19+Tweets&btnG=
3. Vijay, Tanmay, et al. "Sentiment Analysis on COVID-19 Twitter Data." *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE, 2020, https://ieeexplore.ieee.org/document/9358301
4. Manguri, Kamaran H., Rebaz N. Ramadhan, and Pshko R. Mohammed Amin. "Twitter sentiment analysis on worldwide COVID-19 outbreaks." *Kurdistan Journal of Applied Research* (2020): 54-65. https://scholar.google.co.in/scholar?q=twitter+sentiment+analysis+on+worldwide+covid 19+outbreaks&hl=en&as_sdt=0&as_vis=1&oi=scholart
5. Naseem, Usman, et al. "Covidsenti: A large-scale benchmark Twitter data set for COVID- 19 sentiment analysis." *IEEE Transactions on Computational Social Systems* (2021), https://ieeexplore.ieee.org/document/9340540
6. Khan, Rijwan, et al. "Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data." *J. Critical Rev* 7.9 (2020):
7. Wrycza, Stanisław, and Jacek Maślankowski. "Social media users’ opinions on remote work during the COVID-19 pandemic. Thematic and sentiment analysis." *Information Systems Management* 37.4 (2020):288-297, https://www.tandfonline.com/doi/full/10.1080/10580530.2020.1820631
8. Pokharel, Bishwo Prakash. "Twitter sentiment analysis during covid-19 outbreak in nepal." *Available at SSRN 3624719(2020),* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624719
9. Drias, Habiba H., and Yassine Drias. "Mining Twitter Data on COVID-19 for Sentiment analysis and frequent patterns Discovery." *medRxiv* (2020), https://www.medrxiv.org/content/10.1101/2020.05.08.20090464v1
10. Gupta, Prasoon, et al. "Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter." *IEEE Transactions on Computational Social Systems* (2020), https://ieeexplore.ieee.org/document/9301194
11. Pokharel, Bishwo Prakash. "Twitter sentiment analysis during covid-19 outbreak in nepal." *Available at SSRN 3624719*(2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624719
12. Raheja, Supriya, and Anjani Asthana. "Sentimental Analysis of Twitter Comments on Covid-19." *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, https://ieeexplore.ieee.org/document/

9377048

13. Ghasiya, Piyush, and Koji Okamura. "Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach." *IEEE Access* 9 (2021): 36645-36656, https://ieeexplore.ieee.org/document/9366469
14. Machuca, Cristian R., Cristian Gallardo, and Renato M. Toasa. "Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach." *Journal of Physics: Conference Series*. Vol. 1828. No. 1. IOP Publishing, 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1828/1/012104
15. Hussain, Amir, and Aziz Sheikh. "Opportunities for artificial intelligence–enabled social media analysis of public attitudes toward Covid-19 vaccines." *NEJM Catalyst Innovations in Care Delivery* 2.1 (2021). https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0649
16. 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* 18.11 (2021): 5993
17. Rehman, Amir, et al. "COVID-19 Detection Empowered with Machine Learning and Deep Learning Techniques: A Systematic Review." *Applied Sciences* 11.8(2021):3414, https://www.mdpi.com/2076-3417/11/8/3414