IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis on COVID-19 Twitter Data

Justin Seby

Abstract: ***The COVID-19 pandemic has led to the dramatic destruction of human life worldwide and presents an unprecedented hurdle to public health, livelyhood and the world of work. The social and economical disruption caused by the pandemic is devastating. The outbreak first appeared in to spotlight in December 2019 in Wuhan, China. This was declared as a pandemic by the World Health Organization on 11th March 2020. This article analyses how people react to a pandemic outbreak, how much they are aware of the disease and its symptoms, what preventive measures they are taking, whether people are following government guidelines, etc. The experiments have been conducted on the collected data related to COVID-19 tweets from all over the world. This study was conducted to understand how people from different infected countries cope with the situation.***

I. INTRODUCTION

The COVID-19 is the most adverse global public health, economic, social, and socioeconomic stress since the Second World War. It affects all countries and challenges the economy globally, particularly the global supply chain [1]. Given the crucial role of social media as a communication platform, the raw data of emotion is readily available on these different platforms. Social media like Twitter, Facebook has become overloaded with content associated with COVID-19. The impact that social media significantly influence the public as well the private sectors. Public sentiment is a critical factor in deciding on specific services, including airline services, disproportionate to the actual public health need. The Spatio-temporal variability in the discussions on social media, specifically Twitter, is often not in line with the realistic intensity of the outbreak [2].

Sharing short messages to an audience on any social media, called microblogging. This allows people to express their emotions without imposing unwanted communication on someone who might feel obligated to respond. Thus, It can be taken as raw data primarily to extract peoples opinions for any important matter. These data are readily available for the public domain. Different use of social network sites like Twitter makes the process of information sharing significantly faster. Without any surprise, COVID-19 has been one of the tweeted topics on Twitter during these pandemic years. Since almost every country has adopted quarantine measures, various Social media sites like Twitter becomes very popular. Twitter data helps expose public feelings about exciting issues and accurate knowledge of emerging pandemics. In the ongoing COVID-19 pandemic, different government agencies worldwide use Twitter as a necessary means of contact to constantly exchange policy updates and news related to COVID-19 with the general public [3].

In this work, We considered tweets collected from worldwide during the pandemic, and we aim to capture the people's feelings and opinions within the timeframe. To achieve this goal, We defined different machine learning models like s like logistic regression (LR) and support vector machines (SVM) models. Using these models, the study tries to answer the following research questions (RQ).

RQ1: What are the most frequent hashtags used in the positive and negative polarity tweets?
RQ2: What is the effect of COVID-19 on people?
RQ3: How can we achieve extracting people emotions using the ML algorithms?

The tweets posted in English have been considered for the sentiment analysis, which helps understand the problem statement. The collected data will be preprocessed and applied with machine learning algorithms to perform the sentiment analysis.

II. RELATED WORK

For the last ten years, interest in mining sentiment and opinions in the text have multiplied due to the significant increase in the availability of documents and messages expressing personal opinions. Most importantly, Twitter data sentimental analysis has been used to predict or analyze different domains ranging from public opinion to the stock market, politics, and Science. This section mainly covers some research on sentiment analysis using Twitter and other social media. A study has been conducted on English and Arabic Twitter datasets, aiming to investigate the impact of the COVID-19 pandemic using Twitter sentiment analysis through different approaches like k-means clustering and Mini-Batch k-means clustering approaches[4].In the results, The sentiment analysis shows that most tweets are neural in the USA, Australia, Nigeria, Canada, and England. However, both Italy and India have the majority of tweets as positive tweets. This shows that the people of Italy and India are more optimistic than other countries towards the pandemic. Similarly, the k-means cluster in the English dataset shows some patterns where cluster 1 is about COVID-19 pandemic procedures and cluster 3 is about motivating the health workers. On the other hand, the Arabic dataset shows a pattern in the clusters

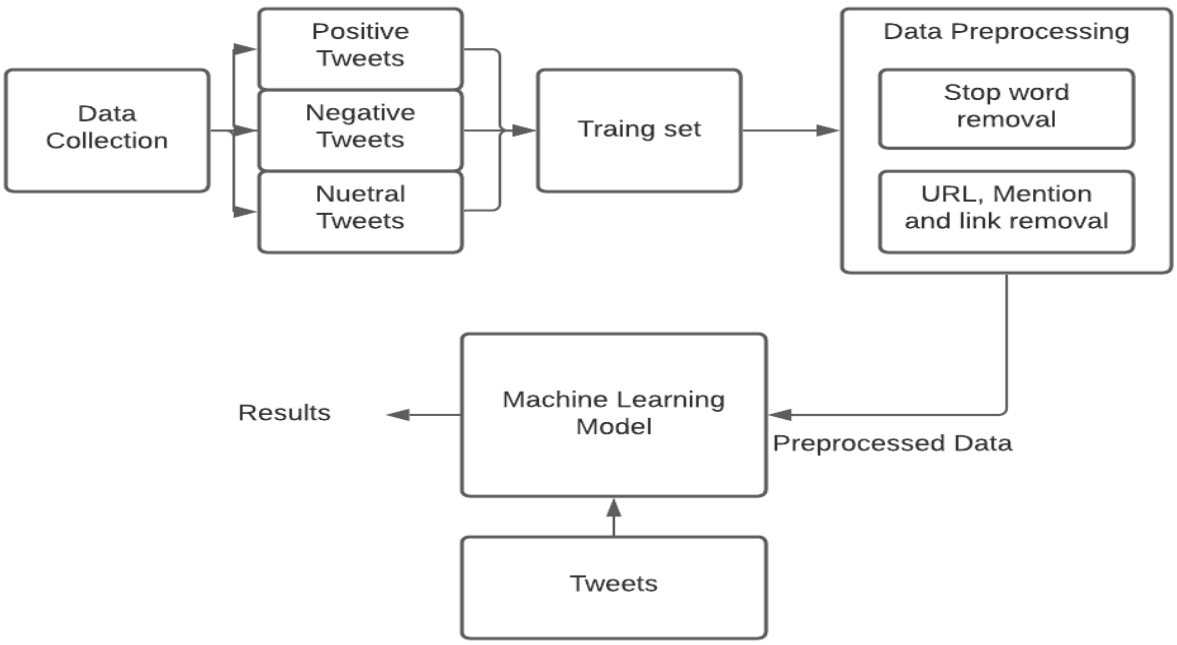
where cluster 1 (with cities' names), 2 (watching media), 3 (about religion), 4 (about coronavirus news), and 8 (psychological group of the word). Another finding is that the mini-batch k-means model required less time to build than k-means with a slight difference in the performance. The limitation of this paper is the usage of one PC with limited memory. Cloud computing such as Amazon Web Services (AWS) can help to apply various clustering methods. Also, there are many spam tweets in Arabic with commercial tweets as advertisements that influence the quality of the data and take more time in cleaning. The lack of Arabic libraries in Python and the limited usage in R makes it exceedingly difficult to deal with the Arabic dataset. Other techniques can be applied in the future to get a better result.

Another study focus on finding people with depression from the Twitter data using machine learning techniques. Social networks have been developed as an excellent point for users to communicate their thoughts and share their photos and videos reflecting their moods, feelings and sentiments. This creates an opportunity to analyze social media data for users' emotions and beliefs to investigate their moods and attitudes when communicating via these online tools[5]. They studied three factors (emotional process, temporal process, linguistic style) and trained a model to utilize each type of factor independently and jointly. For this research, the group analyzed 7146 depressive indicative Facebook comments to identify the most influential time. From the result, we can conclude 54.77% of depression indicate Facebook users communicate with their friends from midnight to midday and 45.22% from noon to midnight.

In the context of Twitter, LI Bing and the team did a similar study on Sentiment Analysis in Twitter Data for Predicting Company's stock price movements[6]. During the research, they have tried to answers the following questions. Can public sentiment from Twitter be analyzed to predict the trends of the stock price of a particular company? If so, is it true that the stock price of one company is more predictable than that of another company? Is there a specific kind of company whose stock prices are more predictable based on analyzing public sentiments reflected in Twitter data?. This is achieved by extracting ambiguous textual tweet data through NLP techniques to define public sentiment, then using a data mining technique to discover patterns between public sentiment and actual stock price movements. This research has two significant practical implications. Firstly, the proposed algorithms have a better prediction accuracy in specific domains such as IT and media. Secondly, the study indicates the proposed algorithms have comparatively better accuracy in using the current tweet's sentiment to predict the stock price three days later. This knowledge informs that a three-day interval is the best period to find and analyze.

III. RESEARCH METHODOLOGY

In this section, The main components of this research are discussed, including data collection and preprocessing, development of machine learning models. Figure 1 shows the proposed model. This study mainly focussed on the Twitter data on COVID-19 due to the fact that the targeted period is COVID time. The availability of raw emotions of the public makes Twitter data best for this study. The tweets were collected from Twitter users all over the world. The frequently repeated hashtags were corona, coronavirus, social distancing etc



Data preprocessing is the process of cleaning the data and convert that into a usable format. Data preprocessing includes various steps like handle missing values, stop word removal, URL removal etc. Since the dataset is huge, manual filling is quite impossible. Popular techniques for handling missing values are replacing them with NA, mean average of all the available values, or in case of non-normal distribution, replacing them with the median value. While using regression or decision tree algorithms, the missing value can be replaced by the most probable value. In some cases, removing the entire row if one attribute contains a null value is also popular. Here, All the missing values are replaced by a dummy value(which is zero). Stop word removal is the process of removing all the words that don't contribute any meaning to the sentence, such as pronouns, conjunctions etc. There are established stopword lists that we can easily import and use. Here in this study, we are using the NLTP library contains a stop word list. At the end of the process, A list returned that doesn't have any words with little to no meaning. The next step of the data preprocessing was to remove all the

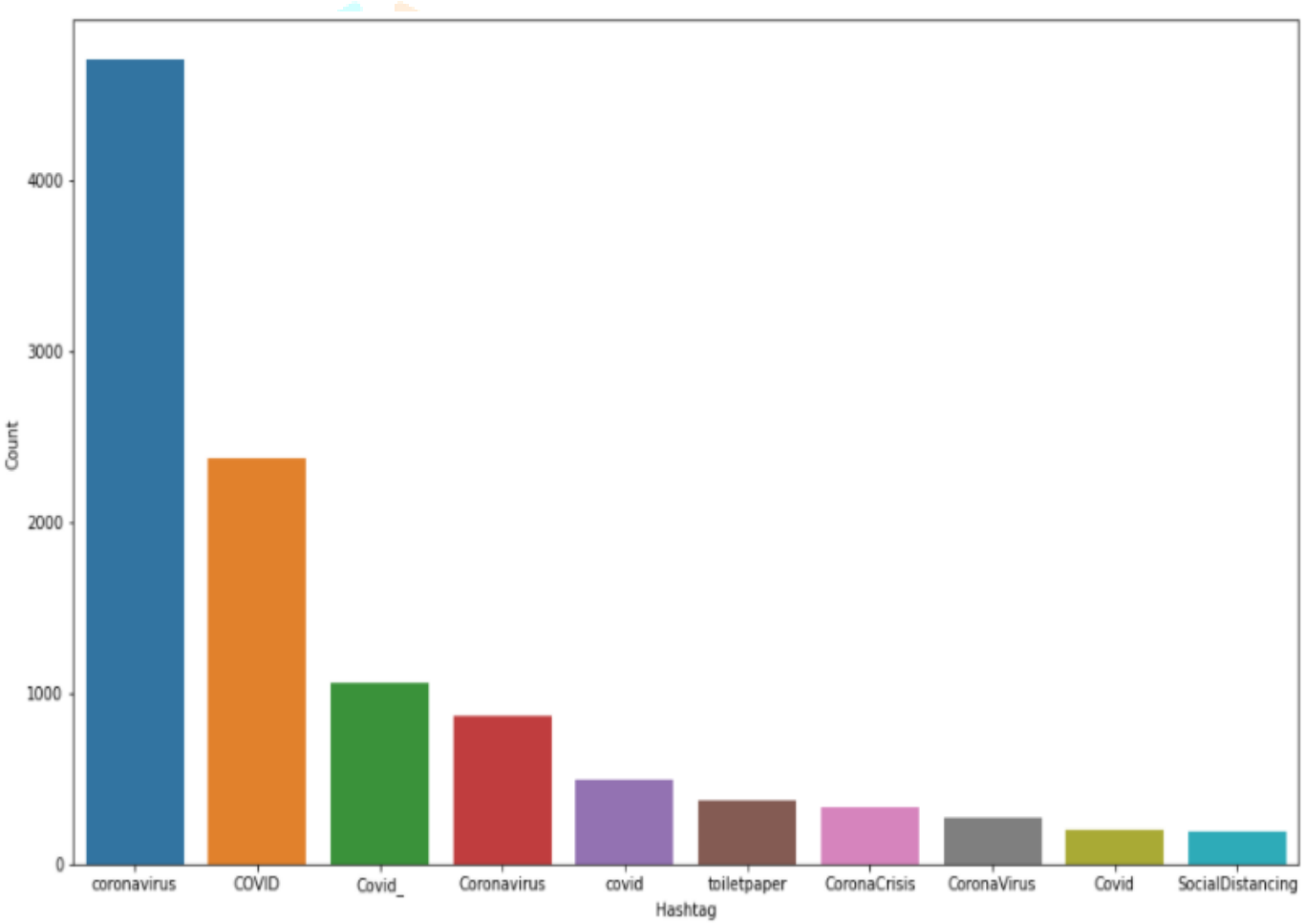
URLs, mentions, digits and HTMKL tags. These properties don't necessarily contribute anything to the meaning of the sentence. Hence, the removal of these sentences doesn't affect our overall objective..

3.2 Machine learning model

The next step after preprocessing and feature extraction were to build a machine learning model. After preparing the Model, we feed the previously prepared dataset into the two ML models: Logistic regression and Support vector machines. Logistic regression otherways known as logit regression, the maximum-entropy classification or the log-linear classifier. In this Model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function. Support vector machines (SVMs) are a set of supervised learning techniques used for classification, regression and outliers detection. It's one of the popular Machine learning models today. There are many advantages to using the SVM model, including being very effective in high dimensional spaces and producing accurate results even if the number of dimensions is greater than the number of samples. SVM is also considered very versatile.

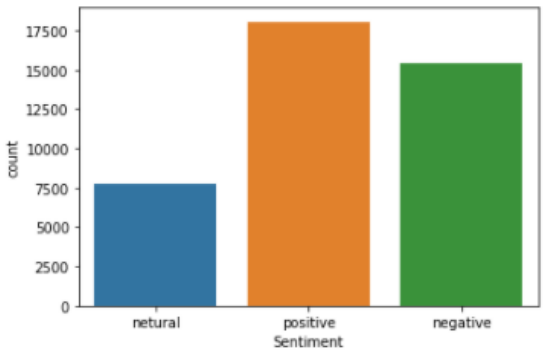
IV. RESULTS AND DISCUSSION

We have used the Twitter data, which is publically available. These readily available datasets feed into the different machine learning models after preprocessing. The machine learning model includes support vector machine and logistic regression model. Based on the results, LR Shows better results than the support vector machine-based model. The support vector machines and Logistic regression model gives an accuracy of 76%, 79%, respectively. Word Cloud is a type of data visualization technique widely used to analyze data from social network websites. In this method, the represented text data indicates its frequency or importance. The frequently used hashtags are also found using ML libraries. The top 10 most used hashtags include Covid, Covid-19, Stayathome etc. The detailed list of the top 15 most used hashtags is displayed in diagram 2.



V. CONCLUSION.

In conclusion, we were able to design a machine learning model for sentimental analysis using this approach. The model has an accuracy of 80%. Two main machine learning techniques used were Support vector machine and Logistic regression. Both the model accuracy was very comparable and differed only in one percentage difference. In the test data, most tweets were positive sentiment followed by negative and neutral in the third



REFERENCES

[1] P. Staszkiewicz, I. Chomiak-Orsa and I. Staszkiewicz, "Dynamics of the COVID-19 Contagion and Mortality: Country Factors, Social Media, and Market Response Evidence From a Global Panel Analysis," in IEEE Access, vol. 8, pp. 106009-106022, 2020, doi: 10.1109/ACCESS.2020.2999614.

[2] Depoux A, Martin S, Karafillakis E, Preet R, Wilder-Smith A, Larson H. The pandemic of social media panic travels faster than the COVID-19 outbreak. J Travel Med. 2020 May 18;27(3):taaa031. doi: 10.1093/jtm/taaa031. PMID: 32125413; PMCID: PMC7107516.

[3] Rufai SR, Bunce C. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. J Public Health (Oxf). 2020 Aug 18;42(3):510-516. doi: 10.1093/pubmed/fdaa049. PMID: 32309854; PMCID: PMC7188178.

[4] M. A. Alanezi and N. M. Hewahi, "Tweets Sentiment Analysis During COVID-19 Pandemic," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-6, doi: 10.1109/ICDABI51230.2020.9325679.

[5] Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst. 2018;6(1):8. Published 2018 Aug 27. doi:10.1007/s13755-018-0046-0

[6] L. Bing, K. C. C. Chan and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements," 2014 IEEE 11th International Conference on e-Business Engineering, 2014, pp. 232-239, doi: 10.1109/ICEBE.2014.47.