Sentiment Prediction on COVID-19 Vaccination Reviews

Nabeel Ahmed
School of Computing
Asia Pacific University of Technology and innovation (APU)
Kuala Lumpur, Malaysia
TP054305@mail.apu.edu.my

Mafas Raheem
School of Computing
Asia Pacific University of Technology and innovation (APU)
Kuala Lumpur, Malaysia
raheem@apu.edu.my

***Abstract*—COVID-19 pandemic has afflicted millions of people worldwide, resulting in three million deaths. Many medical experts and authorities have been striving to combat this pandemic using vaccines. Nowadays, various types of vaccines are available in different countries. Many people around the world have already been vaccinated and many are still receiving. Sharing opinions about vaccines has become a natural phenomenon for the last one plus years where social media platforms play a major role. In this line, people who received the vaccine expressed their sentiments towards the COVID-19 vaccination on Twitter on a large scale. In this study, sentiment predictive models such as Support Vector Machine (SVM), Logistic Regression and Multinomial Naive Bayes were developed with the feature extraction methods like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TFIDF) to predict according to the analysis, the SVM model outperformed with the accuracy of 87.89% (TFIDF) and 87.40% (BoW).**

Keywords—* *COVID-19, vaccination, sentiment analysis, machine learning algorithms, feature extraction methods.

I. Introduction

The ongoing COVID-19 pandemic has affected about 96 million people worldwide, resulting in three million deaths (Worldometers, 2021). Moreover, it has also brought major challenges to global health, economy, food systems and living environments as the virus is highly contagious, spreads quickly and evolves among humans. The symptomatology of the patients was identified as viral pneumonia, which included fever, lethargy, dry cough, and dyspnea (Liu et al., 2020).

It's crucial to have equitable access to safe and effective vaccines to control the COVID-19. Medical experts and authorities from different countries produced various types of vaccines such as Pfizer/BioNTech, AstraZeneca-SK Bio, Serum Institute of India, Janssen, and Moderna that are listed in the World Health Organization's list of emergency vaccinations (World Health Organization, 2021a). But it is not only the vaccine but also the vaccination process that should be effective to end this pandemic. It is important to ensure that vaccines are distributed fairly and equally in all the countries to safeguard their citizens.

People usually criticize any new efforts that are taken in the world in the form of opinions shared on digital platforms. Similarly, types of vaccines and the vaccination processes were also noted as talk of the town on which people shared different opinions. The opinions of people have a high tendency to shape the ideology of others either positively or negatively where the negative influences are usually hard to manage. However, the advancement of Natural Language Processing (NLP) and Machine Learning (ML) techniques offer immense support in predicting the sentiments of the opinions shared by the people on digital platforms. The Twitter platform has been selected as the data source for this study.

Sentiment analysis is one of the prominent text classification applications that identifies and extracts subjective information from the opinions and predicts the sentiments as either positive, negative, or neutral. In sentiment analysis, both NLP and ML techniques got combined to allocate weighted sentiment scores to topics, entities, themes and categories on a phrase or sentence. Social media as a source of textual data is beneficial as huge data possibly be gathered/used for sentiment analysis.

Nowadays most businesses use sentiment analysis to help them understand the social sentiment of their products, brand or service while monitoring online discussions. Therefore, businesses can use social sentiment research to inspire them to be more proactive on social media and communicate with their customers directly.

Twitter is a microblogging platform that allows the user to send and receive short messages known as tweets. These tweets can contain up to 140 characters and can also include links to specific sites and resources (ESRC, 2021). Around 1.3 billion accounts are registered on Twitter with 145 million daily users and 330 million monthly active users. Twitter data is the most accessible source of live, public conversations on the globe, and as a result, it may be a useful tool for gauging customer opinion as people and markets react to product and business decisions (Reddy, 2020). With the help of Twitter, articulated tweets or text data can be analyzed to find the sentiments related to COVID-19 vaccination.

To understand and investigate the impact of the COVID-19 vaccination situation, it is really important to analyze the sentiments of the people globally to make data-driven decisions. However, involving humans to find the sentiments of the people’s opinion could cause any sort of human error or bias. These methods also take a lot of time and are costly too. The accuracy of the result is also difficult to measure and hence also affects the analyzing process which would result in making insufficient decisions at the curial times during the pandemic when people's lives are highly at risk. Moreover, in traditional methods, there is a small amount of audience or scale from which we collect data (like through newspapers, articles, interviews, surveys, forms, etc.) which would affect our analysis and decision-making process as less amount of data may lead to less accurate results.

This research aims to analyze the opinions of people around the globe regarding the COVID-19 vaccination by performing a twitter-based sentiment analysis using a machine learning approach. This could improve the communication and understanding among people which would result in better decision-making. As Twitter is one of the best microblogging platforms with over 187 million daily monetizable active users, it can be used as a suitable platform for gaining a better understanding of public opinion about the COVID-19 vaccination on a large scale. In this project, the NLP concepts and ML models were implemented to perform sentiment analysis in a predictive manner regarding the COVID-19 vaccination around the globe.

II. LITERATURE REVIEW

Gupta & Chen (2020) used sentiment analysis to study the influence of sentiment stated through StockTwits (a microblogging website) on the stock price prediction. To illustrate the effectiveness of the proposed study on stock price prediction, the researchers used nine-month StockTwits data and daily stock data of five firms (Apple, General Electric, Amazon, Target, and Microsoft). They performed sentiment analysis using three machine learning algorithms such as SVM, Naïve Bayes and logistic regression with five featurization techniques such as bigram, trigram, TF-IDF (Term Frequency-Inverse Document Frequency), a bag of words (BoW) and Latent Semantic Analysis (LSA). It was found that for all the five companies the logistic regression model with TF-IDF gave the highest accuracy between 75% and 85%. To determine the impact of sentiments on stock price, correlation analysis and stock price change prediction were also done by the researchers. Finally, to improve the accuracy of stock price movement prediction, sentiment data was integrated with historical stock time-series data. Tables 1,2,3,4 and 5 represents sentiment classification accuracy results for the five firms (Apple = AAPL, Amazon = AMZN, General Electric = GE, Microsoft = MSFT, and Target = TGT) with respect to each combination of methods. Table I – V summarizes the details of the results.

TABLE I. SENTIMENT CLASSIFICATION ACCURACY FOR AAPL

	BoW	Bi-gram	Tri-gram	TF-IDF	LSA
Naïve Bayes	44.9	53.4	54.3	45.4	42.8
Logistic Regression	72.7	73.5	73.8	75.7	74.9
SVM	70.6	70.5	71.7	74.5	74.8

TABLE II. SENTIMENT CLASSIFICATION ACCURACY FOR AMZN

	BoW	Bi-gram	Tri-gram	TF-IDF	LSA
Naïve Bayes	58.4	63.8	63.8	58.9	56.9
Logistic Regression	75.7	76.4	76.8	75.4	75.7
SVM	73.8	74.5	74.5	75.4	75.0

TABLE III. SENTIMENT CLASSIFICATION ACCURACY FOR GE

	BoW	Bi-gram	Tri-gram	TF-IDF	LSA
Naïve Bayes	42.6	55.9	55.9	45.8	46.0
Logistic Regression	73.8	74.2	75.3	76.5	74.8
SVM	73.8	74.5	74.5	75.4	75.0

TABLE IV. SENTIMENT CLASSIFICATION ACCURACY FOR MSFT

	BoW	Bi-gram	Tri-gram	TF-IDF	LSA
Naïve Bayes	63.7	72.2	72.3	64.9	69.3
Logistic Regression	82.6	84.7	85.0	84.8	84.6
SVM	82.8	83.2	84.2	85.6	85.7

TABLE V. SENTIMENT CLASSIFICATION ACCURACY FOR TGT

	BoW	Bi-gram	Tri-gram	TF-IDF	LSA
Naïve Bayes	50.6	58.9	58.8	50.8	54.5
Logistic Regression	73.8	73.3	74.7	74.6	73.6
SVM	71.9	71.8	72.6	74.9	75.7

Villavicencio et al. (2021) developed a sentiment analysis system to understand and analyze the sentiments of people regarding the COVID-19 vaccines in the Philippines. The researchers used the Naïve Bayes model with the TF-IDF technique for classifying English and Filipino language (Tagalog) tweets into negative, positive, and neutral. Moreover, K-fold (K=10) cross-validation was applied to evaluate the performance of the model. The Naïve Bayes model obtained an accuracy of 81.77%.

Machuca et al. (2021) built a binary Logistic Regression model with the TF-IDF technique to perform a twitter-based sentiment analysis on the COVID-19 pandemic. The binary Logistic Regression model gave an accuracy of 78.5% which was considered to be a relatively good performance.

Kim Phung et al. (2021) implemented supervised machine learning methods in mining online customer reviews in the Vietnamese language. In this study, the researchers conducted the training of the model in two ways by using the Hold-Out method and K non-interesting subsets. In the Hold-Out method, the dataset was randomly divided into two subsets based on a common rule: 70% for training and 30% for testing. Whereas in the other method, in each of the trials (out of K times) one subset was used as a test set and the other (K-1) subset was utilized as a training set, where K=5 in this study. The researcher used six machine learning algorithms such as Naïve Bayes (NB), Logistic regression (LR), Neural Network (NN), Random Forest (RF), Decision Tree (DT), and Support Vector Machines (SVM). The results show that the NB model gave the least accuracy and Logistic Regression (LR) obtained the best accuracy rate (about 80%) compared to other models.

Bernal et al. (2021) proposed different machine learning models such as Logistic Regression, Neural Network, Naive Bayes, and Support Vector Machine to analyze the sentiment of tweets about the COVID-19 vaccination campaign in Mexico between late October 2020 and late April 2021. According to their study, the Logistic Regression model with an accuracy of 83.42% was the best to classify tweets into positive and negative categories.

Alhejaili et al. (2021) proposed a machine learning system to analyze the sentiment expressed in Arabic tweets about the COVID-19 vaccination using different machine learning algorithms such as Random Forest, Support Vector Machine, AdaBoost, K-Nearest Neighbors, Gaussian Naïve Bayes, Decision Tree, and Logistic Regression. The dataset was trained using the TF -IDF feature extraction approach, which used unigram, bigram, and trigram features. The Logistic

Regression model gave the highest accuracy of 87.0% with bigram.

Alabid & Katheeth (2021) implemented Support Vector Machine (SVM) and Naïve Bayes (NB) models to perform sentiment analysis on tweets regarding COVID-19 vaccines. Naïve Bayes classifier obtained the accuracy of 80.0% with stemming and stopwords removal.

Table VI summarizes the works of literature gathered from similar studies.

TABLE VI. SUMMARY OF THE LITERATURE REVIEW

Research Paper Author/Date	Research Title	Models Used	Accuracy Results (%)
Gupta & Chen (2020)	Sentiment Analysis for Stock Price Prediction.	Machine learning methods: (Naïve Bayes, SVM, and Logistic Regression) and five featurization techniques:(bag of words, bigram, trigram, TF-IDF, and LSA).	The combination of Logistic Regression and TF-IDF achieved a reasonably high accuracy level, an average of 77.4% for all the five companies.
Villavicencio et al. (2021)	Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes.	Naïve Bayes model with TF-IDF as the feature extraction method.	81.77
Machuca et al. (2021)	Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach.	Binary Logistic Regression model with TF-IDF as the feature extraction method.	78.50
Bernal et al. (2021)	Sentiment Analysis on Twitter About COVID-19 Vaccination in Mexico.	Logistic Regression	83.42
Alhejaili et al. (2021)	Sentiment Analysis of The Covid-19 Vaccine for Arabic Tweets Using Machine Learning.	Logistic Regression with TF -IDF feature extraction approach which uses bigram.	87.0
Alabid & Katheeth (2021)	Sentiment analysis of Twitter posts related to the COVID-19 vaccines.	Naïve Bayes classifier with stemming and removing stopwords applied	80.0
		Support Vector Machine (SVM) with stemming and removing stopwords applied.	78.0
Kim Phung et al. (2021)	A machine learning approach for opinion mining online customer reviews.	Naïve Bayes	49.0
		Logistic Regression	80.0
		Neutral Network	79.0
		Random Forest	69.0
		Decision Tree	71.0
		Support Vector Machine	79.0

III. METHODOLOGY

A. *Introduction*

A methodology is an important framework for planning, organizing, and managing the development of a project to ensure the successful completion of the project. A more suitable methodology was drafted for this work as depicted in a flow chart shown in Fig. 1.

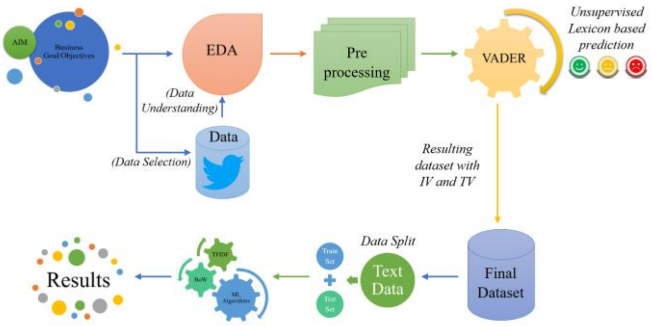

Fig. 1. Methodology flow chart

As shown in Fig. 1, the flow chart comprises several steps providing a cyclical procedure that gives the data mining process an organized approach. The steps include Business goal, Data understanding, Data preparation/preprocessing, VADER operation, Modeling & Evaluation and Results Discussion.

The business goal phase focuses on understanding the objectives and requirements and then turning them into a data mining problem. The goal of this work was to improve the relationship between the organizations and the people regarding COVID-19 vaccination via sentiment analysis of tweets from the general public.

The suitable dataset was acquired from Kaggle (Preda, 2021) and explored in the second stage. It started with the initial data loading process and then progressed through activities to get familiarized with the data, find data quality issues, gain early insights into the data or locate intriguing subsets.

The “vaccination_all_tweets.csv” dataset contains the tweets regarding COVID-19 vaccination on different vaccines such as Sinopharm, Pfizer/BioNTech, Moderna, Oxford/AstraZeneca, Sputnik V, Covaxin and Sinovac (Preda, 2021). This dataset got 228208 records and 16 attributes including id, user_name, user_location, user_description, user_created, user_followers, user_friends, user_favourites, user_verified, date, text, hashtags, source, retweets, favourites, and is_retweet. The most important attribute based on the project’s requirement is the “text” attribute which contains all the tweets of the opinion holders.

A quick overview of the “text” attribute shows that the contained data is valid and got no missing or mismatched data with 226373 unique values.

Furthermore, checking of the statistical measurements for the dataset was also done for a better understanding of data patterns and then determined what data attribute needs to be modified or cleaned. The data preparation includes all processes involved in creating the final dataset from the raw data. The main goal is to align the data with the project or business needs so that the suitable data, cleaning, and extracting attributes from data can be done effectively.

The tweets were cleaned/preprocessed to make them suitable for sentiment analysis. The text cleaning process is very important as it improves the data quality and also helps to build machine learning models with better accuracy. The better the performance of the model, the better the quality and reliability of the result.

B. *Data Preprocessing:*

The cleaning of tweets was done using NLTK via a series of operations such as converting text to lower case, removing unwanted characters or symbols such as “@” or “#” in tweets, removing hyperlinks, removing emojis, removing punctuations, removing stop words, performing lemmatization, or stemming (Lexicon normalization). The existence of such untextual features may hamper the polarity of a tweet, thus this step was done carefully.

VADER (rule/lexicon-based) python library was used to predict the sentiments for the tweets. VADER calculates text sentiment using a collection of lexical features that are categorized as positive, negative, or neutral based on their semantic orientation. The resulting dataset was then used as the final dataset for the supervised machine learning model building for sentiment prediction/classification.

C. *Data Partition:*

The resulted dataset was split into two portions such as 70% for training and 30% for testing.

D. *Modelling & Evaluation:*

Model building was the essential work of this project for building the sentiment classification model to predict or classify the Twitter text data into positive, negative, or neutral sentiments. After performing text preprocessing, a cleaned dataset was obtained suitable for sentiment predictive modelling using the machine learning approach. The bag of words (BoW) technique was used for extracting features from the tweets. The bag of words transforms text into a matrix of word occurrences within a document or in other words, converts text data into numerical feature vectors.

A predictive machine learning model was built using the Multi-Nomial Naïve Bayes (MNB) algorithm as it is very scalable and can handle big datasets with ease. Further, different predictive models with different feature extraction methods were built with the idea of choosing the best-performing model. The Bag of words (BoW) technique was used as a feature extraction method when building the MNB model, but for comparison and evaluation purposes the Term Frequency-Inverse Document Frequency (TF-IDF) was implemented as a feature extraction method to build the Support Vector Machine and Logistic Regression model.

TABLE VII. RESULTS OF THE MODELS

Our Proposed Method	Twitter-Based Sentiment Analysis on COVID-19 Vaccination.	Multinomial Naive Bayes	78.55 (TF-IDF) & 76.32 (BoW)
		Support Vector Machine	87.89 (TF-IDF) & 87.40 (BoW)
		Logistic Regression	85.72 (TF-IDF) & 86.46 (BoW)

Based on the comparison of model performances between the models built in this study and the models from the literature study, the SVM model with TF-IDF obtained the highest accuracy of 87.89%. Further, the SVM model obtained a 0.88 value as the micro average precision. The accuracy value of SVM seemed promising and the model can be deployed in an operational environment for the relevant authorities and the general public to get benefitted.

IV. CONCLUSION

Sentiment analysis is one of the important areas of text analytics which can be implemented on various problems. At present, many businesses have adopted the sentiment analysis as the main tool to understand their customers opinion. On the other hand, sentiment analysis is also applied in the social wellbeing aspects among which the sentiment prediction on the Covid-19 vaccination reviews found to be more crucial. Several predictive machine learning models were built in this regard using the tweets among which SVM model with TF-IDF technique obtained the highest accuracy of 87.89% and outperformed and other models. However, further studies/research can be done in the same area with more data to see a possibility of getting more effective model.

REFERENCES

Alabid, N., & Katheeth, Z. (2021). Sentiment analysis of Twitter posts related to the COVID-19 vaccines. *Indonesian Journal of Electrical Engineering And Computer Science*, 24(3), 1727-1734.

Alhejaili, R., Alhazmi, E., Alsaeedi, A., & Yafooz, W. (2021). Sentiment Analysis of The Covid-19 Vaccine For Arabic Tweets Using Machine Learning. *9th International Conference On Reliability, Infocom Technologies And Optimization (Trends And Future Directions) (ICRITO)*, 1-5. https://doi.org/10.1109/icrito51393.2021.9596517

Athar, A. (2021). *Textblob vs Vader For Sentiment Analysis in Python*. https://www.analyticsvidhya.com/blog/2021/01/sentiment-analysis-vader-or-textblob/

Bernal, C., Bernal, M., Noguera, A., Ponce, H., & Avalos-Gauna, E. (2021). Sentiment Analysis on Twitter About COVID-19 Vaccination in Mexico. *Advances In Soft Computing*, 13068, 96-107. https://doi.org/10.1007/978-3-030-89820-5_8

ESRC. (2021). *What is Twitter and why should you use it? - Economic and Social Research Council*. https://esrc.ukri.org/research/impact-toolkit/social-media/twitter/what-is-twitter/.

Gupta, R., & Chen, M. (2020). Sentiment Analysis for Stock Price Prediction. *2020 IEEE Conference On Multimedia Information Processing And Retrieval (MIPR)*, 1, 213-218. https://doi.org/10.1109/mipr49039.2020.00051

Kim Phung, T., An Te, N., & Thi Thu Ha, T. (2021). A machine learning approach for opinion mining online customer reviews. *2021 21st ACIS International Winter Conference On Software Engineering, Artificial Intelligence, Networking And Parallel/Distributed Computing (SNPD-Winter).* https://doi.org/10.1109/snpdwinter52325.2021.00059

Liu, Y., Kuo, R., & Shih, S. (2020). COVID-19: The first documented coronavirus pandemic in history. *Biomedical Journal*, 43(4), 328-333. https://doi.org/10.1016/j.bj.2020.04.007

Machuca, C., Gallardo, C., & Toasa, R. (2021). Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach. *Journal Of Physics: Conference Series*, 1828(1), 012104. https://doi.org/10.1088/1742-6596/1828/1/012104

Preda, G. (2021). *COVID-19 All Vaccines Tweets*. https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets.

Reddy, R. (2020). *Sentiment Analysis of Twitter Data. Engineering Education (EngEd) Program | Section*. https://www.section.io/engineering-education/sentiment-analysis/.

Villavicencio, C., Macrohon, J., Inbaraj, X., Jeng, J., & Hsieh, J. (2021). Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes. *Information*, 12(5), 204. https://doi.org/10.3390/info12050204

World Health Organization. (2021a). *WHO lists additional COVID-19 vaccine for emergency use and issues interim policy recommendations*. https://www.who.int/news/item/07-05-2021-who-lists-additional-COVID-19-vaccine-for-emergency-use-and-issues-interim-policy-recommendations.

Worldometers. (2021). *COVID Live Update: 207,337,257 Cases and 4,364,041 Deaths from the Coronavirus - Worldometer*. https://www.worldometers.info/coronavirus/#countries.