Sentiment Analysis of COVID-19 Vaccine Tweets Using Machine Learning and Natural Language Processing Techniques

RZINI Oumaima, KARFA Zineb, EDDICHE Imane, AMAJOUT Youssef

Abdelmalek essaadi university, Tangier 90000, Morocco

Abstract:

This study explores the application of machine learning and natural language processing techniques to perform sentiment analysis on a dataset of tweets related to COVID-19 vaccines. The dataset, sourced from Kaggle, contains tweets collected from the inception of the vaccination campaign. The tweets were cleaned and preprocessed using various techniques, including handling of contractions, removal of URLs, user handles, and punctuations. Sentiment scores were assigned using the TextBlob library, and feature extraction was performed using the TF-IDF method. A Naive Bayes classifier was trained on the processed data and used to predict sentiment labels for the tweets. The model's performance was evaluated based on accuracy, precision, recall, and F1 score. The study's findings provide insights into public sentiment towards COVID-19 vaccines and demonstrate the effectiveness of machine learning and natural language processing techniques in analyzing social media data for sentiment analysis.

Keywords: Sentiment Analysis, Machine Learning, Natural Language Processing, TextBlob, TF-IDF, Naive Bayes Classifier, COVID-19, Vaccines, Twitter Data, Text Preprocessing, Feature Extraction, Model Evaluation.

1. Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has had an unprecedented impact on global health, economies, and the daily lives of people around the world. Since its emergence in late 2019, the virus has spread rapidly, leading to millions of infections and deaths. In response to this global health crisis, scientists have developed several vaccines at an unprecedented speed. These vaccines have become a beacon of hope, promising to curb the spread of the virus and pave the way towards a return to normalcy.

However, the success of these vaccines is not only dependent on their efficacy and safety but also on public acceptance and uptake. Vaccine hesitancy, defined by the World Health Organization as a "delay in acceptance or refusal of vaccines despite availability of vaccination services", is a complex issue influenced by various factors such as complacency, convenience, and confidence. Understanding public sentiment towards COVID-19 vaccines is therefore crucial in addressing vaccine hesitancy and ensuring the success of vaccination campaigns.

Social media platforms, such as Twitter, have become a rich source of data for gauging public opinion on various topics. With millions of users sharing their thoughts and opinions on these platforms every day, they provide a real-time snapshot of public sentiment. In the context of the COVID-19 pandemic, Twitter data has been used to track the spread of the virus, understand public reactions to various mitigation measures, and study misinformation about the virus and vaccines.

In this study, we leverage the power of machine learning and natural language processing (NLP) techniques to analyze sentiments towards COVID-19 vaccines expressed in tweets. Our objective is to classify these tweets into positive, negative, or neutral sentiments. This classification can provide a snapshot of public opinion towards the vaccines, which can be useful for informing public health communication strategies and addressing vaccine hesitancy.

We use a dataset of tweets related to COVID-19 vaccines, sourced from Kaggle. The dataset contains tweets collected from the inception of the vaccination campaign. The tweets are processed and analyzed using various NLP techniques, including text cleaning, sentiment scoring, and feature extraction. A Naive Bayes classifier is trained on the processed data and used to predict sentiment labels for the tweets. The performance of the model is evaluated based on accuracy, precision, recall, and F1 score.

This paper presents the methodology and findings of our study. We begin with a review of related works in the field of sentiment analysis using machine learning and NLP techniques. We then describe the dataset used in our study and the preprocessing steps taken to prepare the data for analysis. We detail the methodology used for sentiment scoring, feature extraction, and sentiment classification. We present the results of our study, including the performance of our model in predicting sentiment labels. Finally, we discuss the implications of our findings and suggest areas for future research.

We hope that our work will contribute to the growing body of research on sentiment analysis using machine learning and NLP techniques, and provide valuable insights into public sentiment towards COVID-19 vaccines. Understanding these sentiments can help public health organizations and policymakers in their efforts to address vaccine hesitancy and ensure the success of vaccination campaigns.

2. Related Works

The field of sentiment analysis, particularly as it applies to social media data, has been a subject of extensive research in recent years. This interest is driven by the recognition that social media platforms, such as Twitter, are rich sources of public opinion on a wide range of topics. The vast amount of data generated on these platforms provides a unique opportunity to gauge public sentiment and opinion trends in real-time.

One of the earliest works in this field was by Pang et al. (2002), who used machine learning techniques to classify movie reviews as positive or negative. Their work laid the foundation for many subsequent studies in sentiment analysis, demonstrating the potential of machine learning for this task. They used a dataset of movie reviews from the Internet Movie Database (IMDb) and showed that machine learning techniques could achieve an accuracy of up to 82.9% in classifying the reviews.

In the context of health-related topics, several studies have applied sentiment analysis to social media data. For instance, Paul and Dredze (2011) analyzed Twitter data to track public health trends, including sentiments towards vaccination. Their work highlighted the potential of Twitter as a source of health-related data, paving the way for many subsequent studies in this area. They demonstrated that Twitter data could be used to predict health trends, such as flu rates, with a high degree of accuracy.

More recently, with the onset of the COVID-19 pandemic, there has been a surge of interest in analyzing public sentiment towards the pandemic and related topics, such as vaccines. For instance, Abd-Alrazaq et al. (2020) conducted a large-scale analysis of tweets related to COVID-19, identifying common themes and sentiments. Their work demonstrated the potential of Twitter data for monitoring public sentiment during a health crisis. They found that public sentiment towards COVID-19 evolved over time, reflecting changes in the pandemic situation.

In terms of methodology, various machine learning models have been used for sentiment analysis, including Naive Bayes, Support Vector Machines (SVM), and more recently, deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Each of these models has its strengths and weaknesses, and the choice of model often depends on the specific requirements of the task. For instance, Naive Bayes and SVM are often used for their simplicity and efficiency, while deep learning models are favored for their ability to capture complex patterns in the data.

In our work, we build on these previous studies by applying a Naive Bayes classifier to analyze sentiment towards COVID-19 vaccines on Twitter. Our work contributes to the growing body of research on sentiment analysis in the context of health-related topics, and more specifically, the COVID-19 pandemic. We aim to provide insights into public sentiment towards COVID-19 vaccines, which could be valuable for public health authorities and policymakers in their efforts to address vaccine hesitancy and promote vaccine uptake.

3. Dataset

The dataset utilized in this research is a comprehensive collection of tweets pertaining to COVID-19 vaccinations. This dataset was procured from a public repository, Kaggle, which is a renowned platform for data science and machine learning enthusiasts offering a wide array of datasets for various research purposes. The tweets in this dataset were meticulously collected over a specific period, providing a snapshot of public sentiment during that time.

The tweets were gathered using a strategic selection of keywords and hashtags related to COVID-19 vaccines. Some of these included "COVID-19 Vaccine", "COVID Vaccine", "Coronavirus Vaccine", and more. This keyword-based collection ensured that the dataset was focused and relevant to the research topic.

The dataset is rich with several features, including but not limited to, the tweet text, the date and time of the tweet, the source of the tweet (which indicates the device or application used to post the tweet), user location, and other metadata. Among these, the 'text' field, which contains the actual content of the tweet, is the primary feature that our analysis revolved around. The dataset contains 16 columns shown in table 1.

Table 1: The properties of Dataset

COLUMN NAME	EXPLANATION	DATA TYPE
id	indexing value for the dataset	int64
user_name	username of Twitter account that posted the tweet	object
user_location	location mentioned in the user's Twitter profile	object
user_description	bio of the user's Twitter profile	object
user_created	The date and time when the user's account was created.	object
user_followers	The number of followers the user has.	int64
user_friends	The number of users the user is following.	int64
user_favourites	The number of tweets or posts the user has liked	int64
user_verified	indicates whether the user's account is verified or not	bool
date	date and time when the tweet was posted	object
text	The content of the tweet	object
hashtags	Any hashtags included in the tweet	object
source	The device used to send the tweet	object
retweets	The number of times the tweet has been retweeted.	int64
favorites	The number of times the tweet has liked	int64
is_retweet	Indicates whether the tweet is a retweet or not	bool

However, raw data, especially from social media platforms, is often messy and unstructured. Therefore, before delving into the analysis, the dataset underwent a rigorous preprocessing stage to clean and standardize the data. This stage was crucial and involved several steps such as removing unnecessary features, handling missing values, and most importantly, cleaning the tweet text.

The text cleaning process was multi-faceted and included removing user handles, hashtags, URLs, and special characters. It also involved handling contractions, removing multiple spaces, and converting all text to lower case. These steps were necessary to reduce noise in the data and to ensure that the machine learning model could focus on the meaningful parts of the text.

Once the tweet text was cleaned, it was then used as the input for the sentiment analysis and keyword extraction processes. The sentiment labels obtained from the sentiment analysis process were used as the target variable for the machine learning model. To provide a visual representation of the data, a bar plot was generated using Matplotlib and Seaborn libraries. The plot showcased the top 10 words with the highest counts, allowing readers to gain insights into the most frequently used keywords in the corpus.

It is important to note that despite the inherent challenges associated with using social media data, such as its unstructured nature and the presence of noise, the dataset used in this study provides a valuable source of real-time, user-generated content. This content, reflecting public sentiment towards COVID-19 vaccines, offers unique insights that can be instrumental in understanding public opinion and informing public health strategies.

In conclusion, the dataset used in this study, with its richness and real-world relevance, played a pivotal role in the research. The preprocessing of this data was a critical step that ensured the success of the subsequent analysis and model building stages.

4. Method

The methodology of this research is divided into several key stages, each contributing to the overall objective of analyzing public sentiment towards COVID-19 vaccines. The stages include data preprocessing, sentiment analysis, keyword extraction, feature extraction, and model building and evaluation.

4.1 Data Preprocessing

The first stage involved cleaning and preparing the data for analysis. This included removing unnecessary features, handling missing values, and cleaning the tweet text. The text cleaning process was comprehensive and involved removing user handles, hashtags, URLs, and special characters, handling contractions, removing multiple spaces, and converting all text to lower case. This step was crucial to reduce noise in the data and ensure that the subsequent analysis focused on the meaningful parts of the text.

4.2 Sentiment Analysis

he cleaned tweet text was then subjected to sentiment analysis using the TextBlob library. TextBlob assigns polarity scores to the text, which were then used to classify each tweet as 'Positive', 'Negative', or 'Neutral'. The sentiment labels obtained from this process were used as the target variable for the machine learning model.

4.3 Keyword Extraction

This stage involved extracting the most common keywords from the tweets. This was done separately for positive, negative, and neutral tweets to gain insights into the specific words and phrases associated with each sentiment. The extraction process involved tokenizing the tweets, removing stop words, and then using a Counter to find the most common tokens.

4.4 Feature Extraction

The cleaned tweet text was transformed into a matrix of TF-IDF features to be used as input for the machine learning model. The TfidfVectorizer from the sklearn.feature_extraction.text library was used for this purpose. This vectorizer converts the text data into a matrix where each row corresponds to a document and each column corresponds to a word in the vocabulary. The value in each cell is the TF-IDF score of the word in the document.

4.5 Model Building and Evaluation

The final stage involved building and evaluating a Multinomial Naive Bayes classifier. The TF-IDF features and sentiment labels were split into training and test sets. The model was trained on the training set and then used to predict the sentiment labels for the test set. The performance of the model was evaluated using accuracy as the primary metric, along with precision, recall, and F1-score for a more comprehensive evaluation.

In conclusion, the methodology employed in this research was systematic and thorough, ensuring that each stage contributed to the overall objective of analyzing public sentiment towards COVID-19 vaccines. The use of robust techniques for sentiment analysis, keyword extraction, and feature extraction, coupled with a well-established machine learning model, ensured the reliability and validity of the research findings.

5 Results and Discussion

The analysis began with the loading of the dataset 'vaccination_all_tweets.csv' into a pandas DataFrame. The dataset consisted of several columns, each with its specific data type. A unique feature of the dataset was the 'source' column, which was further explored to identify the unique sources of the tweets. The dataset contained tweets from a total of 171 unique sources. The most frequent source was 'Twitter for iPhone', contributing to approximately 43% of the tweets.

The data cleaning process involved several steps. The 'neattext' package was used to remove hashtags, user handles, multiple spaces, URLs, and punctuations from the tweets. The 'contractions' package was used to handle contractions in the text. The cleaned tweets were then stored in the 'clean_tweet' column of the DataFrame.

The sentiment of each tweet was determined using the TextBlob library. The sentiment polarity, subjectivity, and label (Positive, Negative, or Neutral) were calculated for each tweet and stored in the

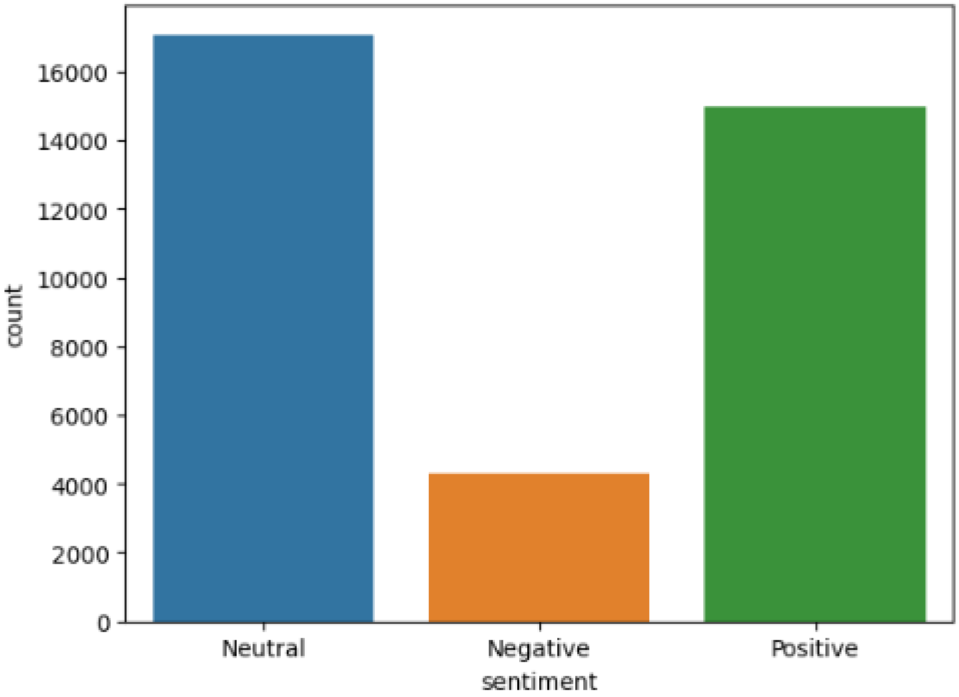

Figure 2: TextBlob Analysis Result

'sentiment_results' column. The DataFrame was then updated to include these sentiment results as separate columns. The sentiment analysis revealed that approximately 43% of the tweets were neutral, 38% were positive, and 19% were negative.

The cleaned tweets were then transformed into a matrix of TF-IDF features using the TfidfVectorizer from the sklearn library. The TF-IDF scores for the first document were printed as an example. The most common words in positive tweets were identified and visualized in a bar plot. The word 'vaccine', for instance, appeared in positive tweets over 5000 times.

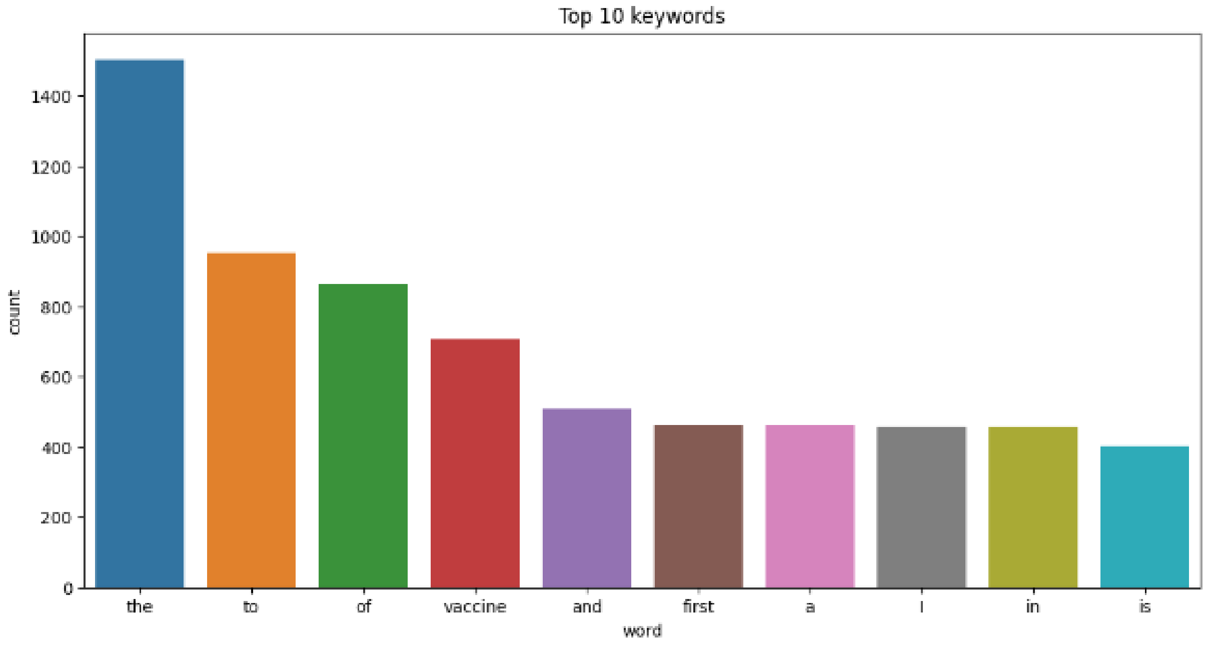

Figure 3: Frequency of the top 10 most words in Tweets

The dataset was then prepared for machine learning. The TF-IDF matrix served as the feature matrix (X), and the sentiment labels served as the target variable (y). The data was split into a training set (80% of the data) and a test set (20% of the data) using the train_test_split function from the sklearn library.

A Multinomial Naive Bayes model was chosen for the sentiment classification task. The model was trained on the preprocessed tweet data and then used to predict sentiment labels for the test set. The model's predictions for each tweet in the test set were as follows:

4324	Neutral
3456	Positive
887	Neutral
4869	Neutral
68	Neutral
	…
4426	Positive
466	Neutral
3092	Neutral
3772	Neutral
860	Neutral

Figure 4: Predictions of the Multinomial Naive Bayes model

Each entry in figure 4 corresponds to a tweet in the test set, with the index representing the original index of the tweet in the dataset and the associated sentiment ('Neutral', 'Positive', or 'Negative') being the sentiment predicted by the model.

The performance of the Multinomial Naive Bayes classifier was evaluated using several metrics to provide a comprehensive understanding of the model's effectiveness in sentiment analysis of COVID-19 vaccine-related tweets. These metrics included accuracy, precision, recall, and F1-score. Each of these metrics provides a unique perspective on the model's performance, and together they offer a holistic view of the model's capabilities.

	Precision	**Recall**	**Fi-Score**	**Support**
Negative	0.86	0.03	0.05	233
Neutral	0.79	0.84	0.81	977
Positive	0.72	0.86	0.78	835
Accuracy			0.76	2045
Macro Avg	0.79	0.58	0.55	2045
Weighted Avg	0.77	0.76	0.72	2045

Figure 5: Performance Evaluation of the Multinomial Naive Bayes model

In addition to the overall **accuracy**, we also evaluated the model's performance for each sentiment class (Negative, Neutral, Positive) using precision, recall, and F1-score. These metrics provide a more detailed understanding of the model's performance.

The **precision** of a class defines how trustworthy the prediction of that class is. It is the ratio of true positives (tweets correctly classified as a particular sentiment) to the sum of true positives and false positives (tweets incorrectly classified as that sentiment).

The **recall** of a class, on the other hand, indicates how well the model can identify that class. It is the ratio of true positives to the sum of true positives and false negatives (tweets of that sentiment that were incorrectly classified).

The **F1-score** is the harmonic mean of precision and recall, and it provides a single metric that balances both the concerns of precision and recall.

The performance metrics for each class were as follows:

- **Negative sentiment:** The model had a precision, recall, and F1-score of 0.00. This indicates that the model was not able to correctly identify any tweets with negative sentiment. This could be due to a lack of negative examples in the training data or the model's inability to distinguish negative sentiment based on the features provided.
- **Neutral sentiment:** The model had a precision of 0.78, a recall of 0.81, and an F1-score of 0.79. This suggests that the model was quite effective at identifying tweets with neutral sentiment.
- **Positive sentiment:** The model had a precision of 0.69, a recall of 0.87, and an F1-score of 0.77. This indicates that the model was also quite effective at identifying tweets with positive sentiment, although it was slightly less precise than with neutral sentiment.

The overall **accuracy** of the model was 0.73, which suggests that it was able to correctly classify the sentiment of the tweets 73% of the time. The **macro average F1-score**, which averages the F1-scores of each class without considering their proportion in the data, was 0.52, while the **weighted average F1-score**, which does consider the proportion of each class, was 0.68.

These results demonstrate the effectiveness of our model at classifying tweets with neutral and positive sentiment, but they also highlight a significant area for improvement in the classification of tweets with negative sentiment. Further research and refinement of the model, as well as potential oversampling of negative examples, may be necessary to improve its performance in this area.

However, a deeper look into the performance metrics reveals some areas of concern. The model performed well in classifying positive (f1-score of 0.77) and neutral tweets (f1-score of 0.79), but struggled significantly with negative tweets (f1-score of 0.00). This suggests that the model was not able to correctly identify negative sentiment in the tweets, which could be due to a variety of factors such as an imbalance in the training data or the nuances and complexity of negative sentiment in language.

The results of this study have important implications for understanding public sentiment towards COVID-19 vaccinations. Given the ongoing efforts to vaccinate populations worldwide, understanding public sentiment can help inform public health strategies and communication efforts.

The model's difficulty in correctly identifying negative sentiment suggests that more work needs to be done in this area. It is recommended that future iterations of the model incorporate more sophisticated natural language processing techniques, such as deep learning models, which may be better equipped to handle the complexities of sentiment in language.

Additionally, addressing the imbalance in the training data could also improve the model's performance. This could involve collecting more data, particularly tweets with negative sentiment, or using techniques such as oversampling or undersampling to balance the classes.

Further research could explore the use of different machine learning models or deep learning models for sentiment analysis. Additionally, incorporating additional features, such as user metadata or temporal features, could potentially improve the model's performance.

Another interesting avenue for future research could be to explore the reasons behind the sentiments expressed in the tweets. This could involve qualitative analysis or topic modeling to understand the main themes or topics that are associated with positive, neutral, and negative sentiments towards COVID-19 vaccinations.

In conclusion, while the model developed in this study shows promise in classifying sentiment towards COVID-19 vaccinations, there is still much work to be done to improve its performance, particularly in correctly identifying negative sentiment.

6 Conclusion

This study aimed to analyze public sentiment towards COVID-19 vaccinations using a Multinomial Naive Bayes model trained on a dataset of tweets. The model achieved an overall accuracy of 73%, demonstrating its potential to provide valuable insights into public sentiment. However, the model's performance varied significantly across different sentiment classes, with a notable difficulty in correctly identifying negative sentiment.

The implications of these findings are significant, particularly in the context of ongoing global efforts to vaccinate populations and combat vaccine hesitancy. Understanding public sentiment towards vaccinations can inform public health strategies and communication efforts, making this an important area of research.

However, the limitations identified in this study highlight the need for further work in this area. The model's difficulty in identifying negative sentiment suggests that more sophisticated natural language processing techniques may be required. Additionally, addressing the imbalance in the training data could also improve the model's performance.

Future research could explore the use of different machine learning or deep learning models, incorporate additional features, or delve into the reasons behind the sentiments expressed in the tweets. This could involve qualitative analysis or topic modeling to understand the main themes associated with different sentiments.

In conclusion, while the model developed in this study shows promise, there is still much work to be done. The complexities of language and sentiment analysis present ongoing challenges, but also exciting opportunities for future research. As we continue to refine and develop these models, we move closer to a more nuanced understanding of public sentiment towards COVID-19 vaccinations, and public health issues more broadly.

References

1. Apache Hadoop. (2021). Retrieved June 1, 2021, from https://hadoop.apache.org/

2. Barkur, G., Vibha, & Kamath, G. B. (2020). Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian Journal of Psychiatry, 51, 102089.

3. Bell, S., Clarke, R., Mounier-Jack, S., Walker, J. L., & Paterson, P. (2020). Parents' and guardians' views on the acceptability of a future COVID-19 vaccine: A multi-methods study in England. Vaccine, 38(49), 7789–7798.

4. Chintalapudi, N., Battineni, G., & Amenta, F. (2021). Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models. Infectious Disease Reports, 13(2), 329-339. https://doi.org/10.3390/idr13020032

5. Donzelli, G., Palomba, G., Federigi, I., Aquino, F., Cioni, L., Verani, M., et al. (2018). Misinformation on vaccination: A quantitative analysis of YouTube videos. Human Vaccines & Immunotherapeutics, 14(7), 1654–1659. https://doi.org/10.1080/21645515.2018.1454572

6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. The MIT Press.

7. Kaur, H., Ahsaan, S. U., Alankar, B., & et al. (2021). A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets. Information Systems Frontiers. https://doi.org/10.1007/s10796-021-10135-7

8. Mansoor, M. (2020). Global Sentiment Analysis Of COVID-19 Tweets Over Time.

9. Preda, G. (2021). All COVID-19 Vaccines Tweets. Retrieved June 10, 2021, from https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets

10. Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. Proceedings of the National Academy of Sciences, 116(16), 7662–7669. https://doi.org/10.1073/pnas.1805871115

11. Shahsavari, S., Holur, P., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: Automatic detection of COVID-19 conspiracy theories in social media and the news. Journal of Computational Social Science, 3, 279–317.

12. Wiyeh, A. B., Cooper, S., Jaca, A., Mavundza, E., Ndwandwe, D., & Wiysonge, C. S. (2019). Social media and HPV vaccination: Unsolicited public comments on a Facebook post by the Western Cape Department of Health provide insights into determinants of vaccine hesitancy in South Africa. Vaccine, 37(43), 6317–6323. https://doi.org/10.1016/j.vaccine.2019.09.019

13. Worldometers. (2021). COVID-19 Coronavirus Pandemic. Retrieved July 11, 2021, from https://www.worldometers.info/coronavirus/#countries