010 Element – 2022 MOD007893 TRI2 F01CAM

Applied Data Analysis & Research Methodology

Final Assignment

Sentiment Analysis on Vaccination Tweets

Thi Nhu Ngoc VO (Nellie VO) SID 2179088

Please click here to be directed to the published online Github repository.

Cambridge, United Kingdom – April 2022

Contents

Abbreviations

API	Application Programming Interface
BI	Business Intelligence
COVID-19	Coronavirus Disease 2019
DPIA	Data Protection Impact Assessment
EDA	Exploratory Data Analysis
GDPR	General Data Protection Regulation
HTTPS	Hypertext Transfer Protocol Secure
LR	Logistic Regression
ML	Machine Learning
MNB	Multinomial Nave Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PII	Personal Identifiable Information
URLs	Uniform Resource Locator
VADER	Valence Aware Dictionary for Sentiment Reasoning

List of Figures

Abstract

The SARS-CoV-2 coronavirus disease (COVID-19) pandemic continues to have an influence on the worldwide population's health and well-being. Controlling the spread of COVID-19 requires vaccination of the whole populace in order to avoid an eventual pandemic outbreak. However, there is some scepticism among the public about vaccinations. In this study, we analyse public tweets from Twitter on COVID-19 vaccines in order to determine the user's perspective on the vaccination. NLTK VADER and TextBlob were used to categorise tweets as positive, neutral, or negative.

Project purpose: to analyze sentiment and opinions expressed in COVID-19 vaccine-related tweets in order to provide insights and recommendations for stakeholders.

Project motivation: to understand how individuals are responding to COVID-19 vaccination programs on social media and to provide valuable insights to support public health efforts.

Design, approach, and method: The project utilizes a dataset of 11,020 tweets. The collected data is analyzed by using Jupyter Notebook where collection, EDA, preprocessing, model training, deployment, and evaluation were implemented. The model is then visualised with Power BI.

Main findings: Both the NLTK VADER and TextBlob analyzers produced relatively similar results in terms of sentiment classification. The majority of tweets were classified as neutral, followed by positive and then negative. The percentage of negative tweets was higher with the VADER analyzer compared to TextBlob.

Practical implication: The project provides practical insights into the public perception of vaccination efforts, helping public health officials identify areas where more education or outreach may be needed to improve vaccination rates and combat vaccine hesitancy.

Keywords: Twitter Sentiment Analysis, Machine Learning, COVID-19 Vaccination, VADER, TextBlob

1. Introduction

In recent times, the world has witnessed an unprecedented surge in the use of social media platforms such as Twitter for information sharing and gathering. One topic that has gained considerable attention on Twitter is the COVID-19 vaccination. While many users express their support for vaccination, others voice concerns or opposition. This presents an opportunity for sentiment analysis, which can help understand public opinion and monitor the spread of misinformation related to COVID-19 vaccination on social media. By analyzing tweets related to COVID-19 vaccination, we can gain insights into the sentiment of the public towards the vaccine, and identify areas that need more attention or clarification. This use case is justified by the need to analyze media content and discover new insights that can aid public health efforts in promoting vaccination and addressing concerns or misinformation. In this context, this project aims to develop a sentiment analysis model for Twitter data related to COVID-19 vaccination, with the goal of helping public health officials and policymakers make data-driven decisions.

2. Literature Review

Sentiment analysis, also referred to as "opinion mining" or "emotion artificial intelligence," refers to the methodical identification, extrication, evaluation, and analysis of emotional states and subjective information using natural language processing (NLP), text mining, computational linguistics, and bio measurements (Alsaeedi and Khan, 2019). Sentiment research often examines the viewpoint expressed in client materials, such as online polls, reviews, and social media platforms. Sentiment analysis differentiates and categorises the author's viewpoint expressed in a given segment of text concerning the premise's foundational topic. The primary purpose of sentiment analysis is to establish the rate of polarity by which the author's tone in a corpus may be classified as positive, negative, or neutral (Praveen, Ittamalla and Deepak, 2021).

There is currently a vast amount of opinionated material available on the Web through social media. A lot of research would not have been possible without this information. Unsurprisingly, sentiment analysis's inception and rapid growth followed those of social media. In fact, sentiment analysis is currently at the core of social media research (Liu, 2012). The way individuals get information from people or organisations they are interested in has been shaped and transformed by TWITTER, a well-liked micro-blogging service throughout the world. Users of Twitter may post tweets, or status updates, to let their followers know what they are thinking, doing, or what is going on in the world. By responding to or reposting another user's tweets, users may also communicate with one another. Twitter has grown to be one of the biggest social networking websites in the world since its founding in 2006 (Li, Dombrowski and Brady, 2018). Because to its numerous uses, mining users' stated attitude polarity in Twitter messages has emerged as a popular research topic in light of the ever-increasing volume of data accessible from Twitter

(Tang *et al.*, 2015). For instance, various algorithms have been created to propose political election plans by evaluating the sentiment polarisation of Twitter users towards political parties and politicians (Wang *et al.*, 2012; Paul *et al.*, 2017). Also, businesses may quickly and efficiently track consumer opinion about their products and brands by employing Twitter sentiment analysis (Bravo-Marquez, Frank and Pfahringer, 2016).

Twitter is a valuable source for data collection and research, offering an organic data source for sociological, political, economic, and analytical analyses. Sentiment analysis, the study of individual and group reactions to a specific topic, can now be automated through machine learning models, providing companies with a cost-effective and time-efficient way to gather public opinion about their products and services.

3. Project Environment

3.1. Project setting

Jupyter Notebook imports vaccination_tweets-1.csv for prediction and analysis. Python runs queries, experiments and analyses. Natural Language Tool Kit (NLTK) Valence Aware Dictionary and Sentiment Reasoner (VADER) and TextBlob are examined.

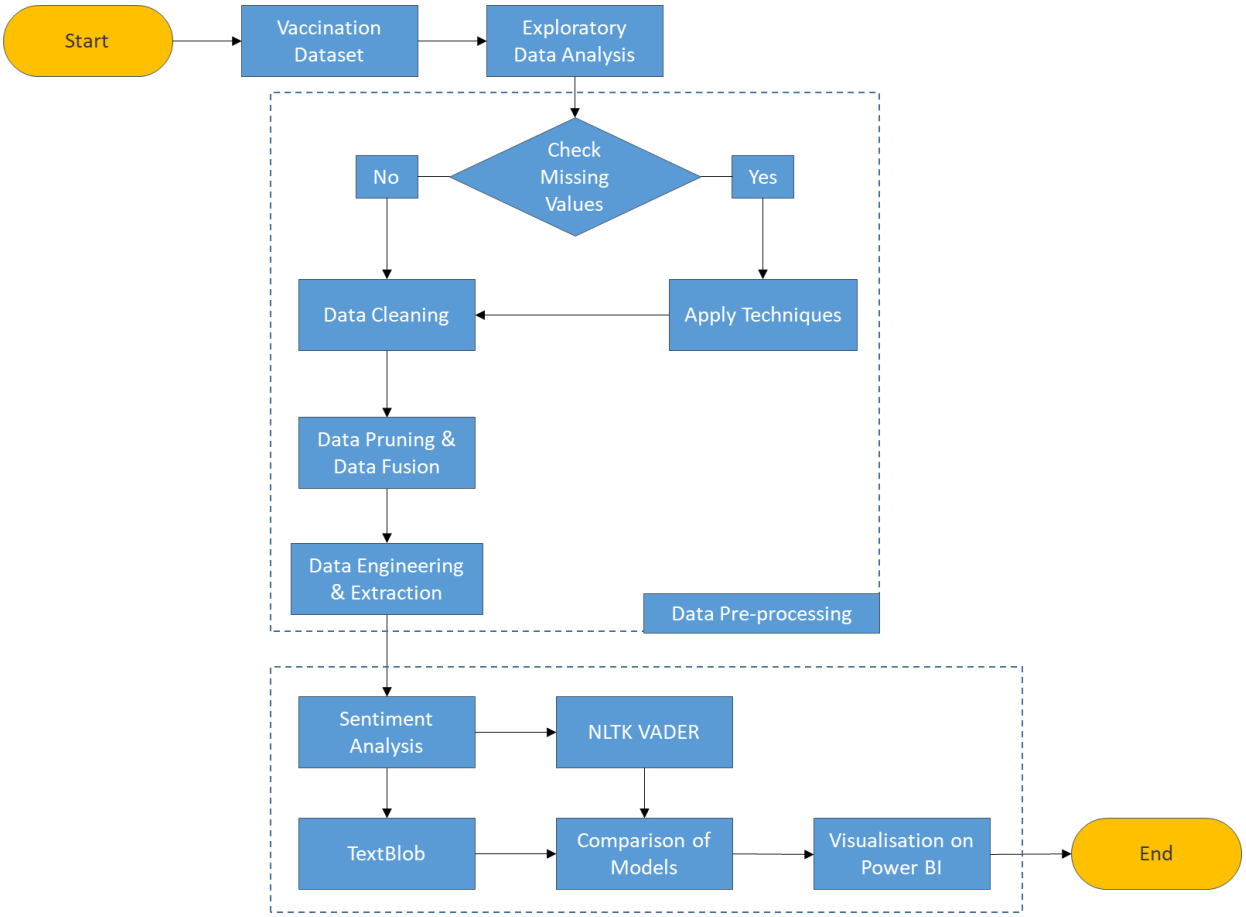
3.2. Dataset

The text_dataset contains 11,020 tweets from the vaccination_tweets-1 dataset. It includes the date, text, number of retweets and favorites of each tweet for predicting and analyzing sentiment towards vaccination on Twitter.

4. Data Pipeline Attributes

Figure 1 depicts the proposed approach's methodology. Details are explained in this part’s sub-sections.

Figure 1: Data pipeline flow chart



4.1. Data collection

4.1.1. Data gathering requirement

Data is gathered in the form of tweets containing relevant keywords related to COVID-19 vaccination. Subjected to analysis is the text data from these tweets to extract sentiment polarity and explore the patterns and trends in the public opinion regarding the covid vaccination.

4.1.2. Data collection

- Tweepy is a popular Python library for accessing the Twitter API. To use Tweepy, developer access to the Twitter API is required, which involves creating a developer account and obtaining API keys. Once API keys are obtained with relevant authorisation, Tweepy can be used to stream and search for tweets related to COVID-19 vaccination. Tweets can be filtered based on specific keywords, hashtags, or user accounts. Data is synchronized at scheduled intervals.
- Destination is Microsoft Power BI. Power BI Desktop is an on-premises application that can be used to create and publish reports and visualizations to the Power BI Service, which is a cloud-based platform.

- Data ingestion model is batch processing, in which source data is collected periodically and sent to the destination system.

4.2. Importing required libraries and dataset in Jupyter Notebook

- Required libraries include: NumPy, Pandas, Matplotlib, Seaborn, SciKit Learn, Natural Language Tool Kit, TextBlob, etc.
- Dataset mentioned in section 3.2. is imported.

4.3. Exploratory data analysis (EDA)

- EDA is deployed to identify necessary pre-processing steps, showing dataset's column names, datatypes, and statistical description.
- Necessary columns that are retained for processing & analysis are 'date', 'text', 'retweets', 'favorites'.
- Distribution of text length, number of tweets overtime, word cloud, and most frequent words diagrams are plotted.
- The majority of tweets have between 120 and 140 characters.
- The dataset contains stopwords, URLs, etc.

4.4. Data preprocessing

4.4.1. Cleaning, transforming, and reformatting

- Dataset is checked for missing values.
- Text column is transformed into all lowercase letters using str.lower().
- URLs are removed.
- Punctuations are removed.
- Single characters and double spaces are removed.
- Emails are removed.
- Text column is tokenised using RegexpTokenizer().
- Stopwords are removed.
- Emoticon are replaced with positive/negative using Emoticon Lookup Table.
- Stemming is applied via PorterStemmer().
- Lemmatizer is applied via WordNetLemmatizer().

4.4.2. Feature extraction

- Feature extraction is employed via CountVectorizer() to select relevant features from the transformed and pruned data to reduce dimensionality and remove redundancies.

4.5. Sentiment analysis

4.5.1. NLTK VADER

NLTK is a free open-source Python package that includes a number of tools for programming and data classification. Linguists, engineers, students, educators, researchers, and developers who deal with textual data in natural language processing and text analytics will benefit from NLTK (Bird, Klein and Loper, 2009). NLTK makes it simple to access the interfaces of over 50 corpora and lexical resources. It consists of a collection of ext processing libraries for categorization, tokenization, stemming, tagging, parsing, and semantic reasoning (*NLTK :: Natural Language Toolkit*, no date).

VADER is a vocabulary and rule-based sentiment analysis tool that is especially geared to social media sentiments. It is a completely free and open-source utility. VADER additionally takes word order and degree modifiers into account (Hutto and Gilbert, 2014).

SentimentIntensityAnalyzer() is employed to extract polarity_scores(), thereby compound score from the text column. Neutral thresh is 0.05 so that those whose compound score is between -0.05 and 0.05 can be classified as neutral. Scores are then categorised into the sentiments of positive, neutral or negative.

Each sentiment class's frequency is counted and plotted on percentage.

4.5.2. TextBlob

TextBlob is a Python (2 and 3) text processing package. It offers a straightforward API for delving into typical natural language processing (NLP) activities including part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more (Akash, 2021).

TextBlob() is employed to extract sentiment.polarity, which is textblob_score. Neutral thresh is 0.05 so that those whose compound score is between -0.05 and 0.05 can be classified as neutral. Scores are then categorised into the sentiments of positive, neutral or negative.

Each sentiment class's frequency is counted and plotted on percentage.

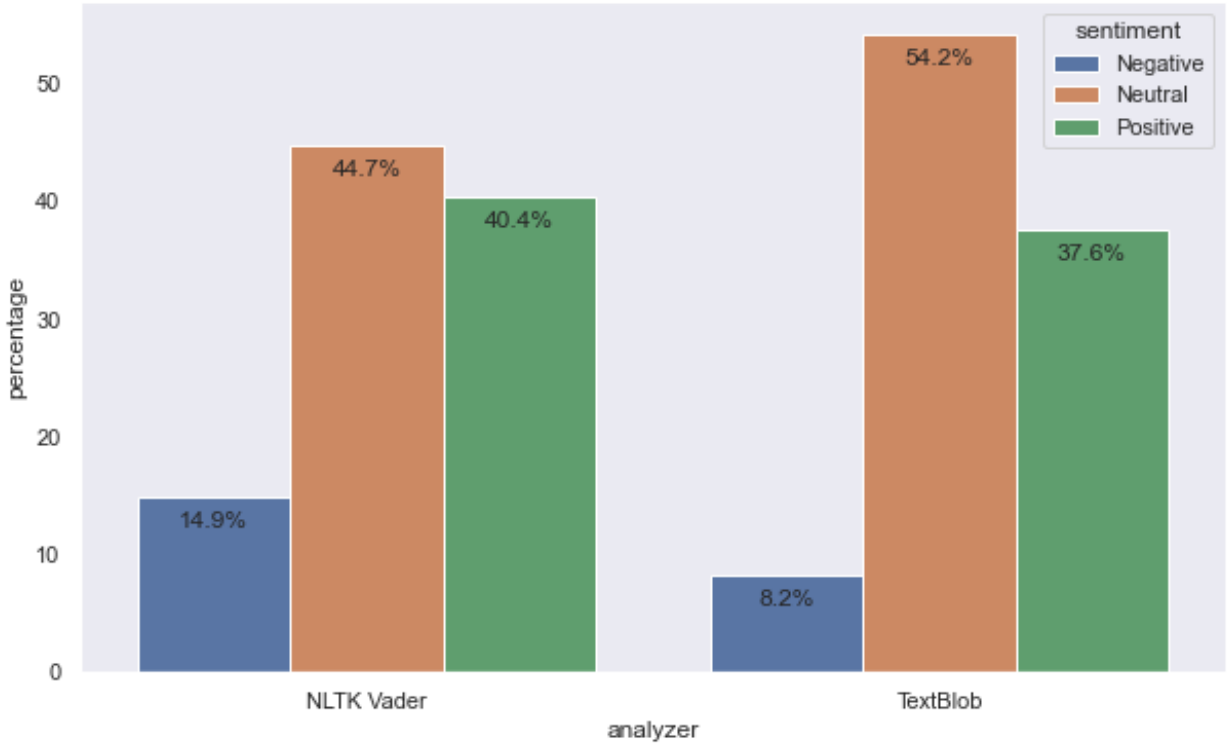
4.6. Model comparison

Sentiment class dataframes from the two models are concatented into one dataframe and a pivot table.

They are compared with a bar graph on percentage of each sentiment class for each analyser and correlation between scores and number of favorites and retweets.

5. Data Visualisation with Power BI & Story Telling Analysis

Figure 2: Comparison on models

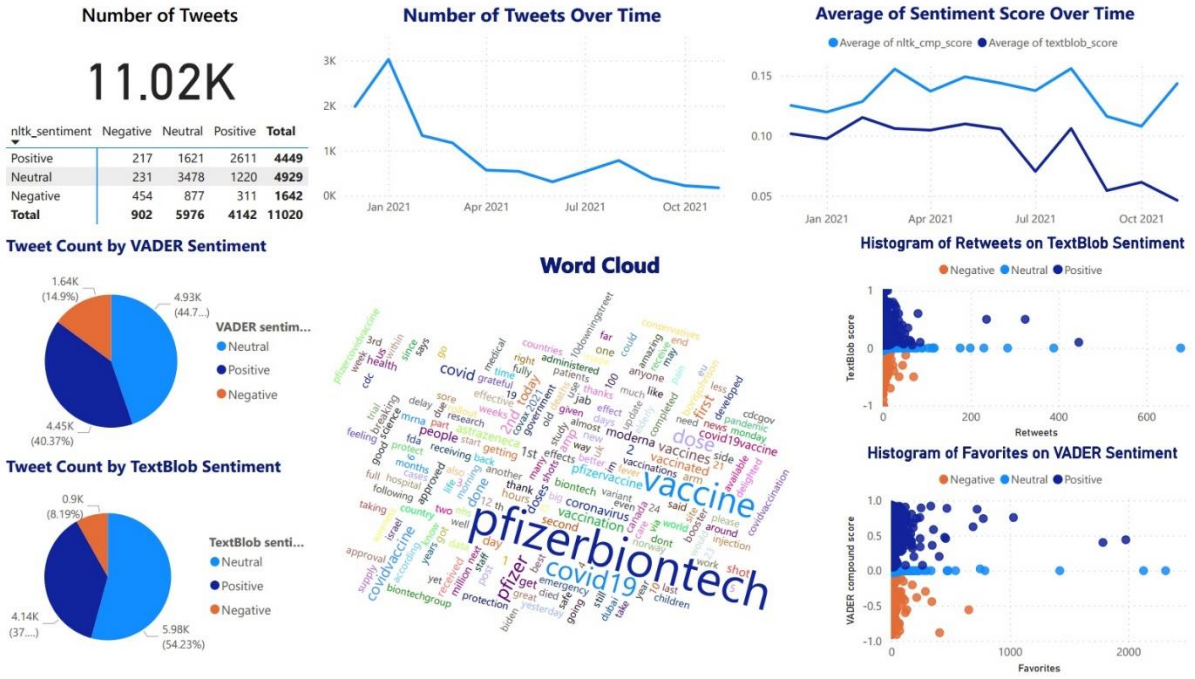


The VADER analyzer has classified 14.9% of the tweets as negative, 44.7% as neutral, and 40.4% as positive. On the other hand, the TextBlob analyzer has classified 8.2% of the tweets as negative, 54.2% as neutral, and 37.6% as positive.

One possible reason for the difference in results could be the fact that the two analyzers use different approaches to sentiment analysis. The VADER analyzer relies on a lexicon of words and phrases that are associated with positive or negative sentiment, while the TextBlob analyzer uses a combination of a rule-based approach and machine learning algorithms.

Another factor that could have influenced the results is the specific dataset that was used for the analysis. The sentiment of tweets related to COVID-19 vaccination may be particularly complex and varied, and this could have affected the accuracy of the sentiment analysis algorithms.

Figure 3: Visualisation with Power BI



After analyzing the Twitter sentiment on COVID-19 vaccination using the VADER and TextBlob models, a detailed visualization using Power BI revealed some interesting insights. The data showed that the number of tweets peaked in January 2021, which is likely due to the increased availability of vaccines at the time. However, since then, the number of tweets has decreased steadily, indicating that the initial excitement surrounding the vaccination campaign may have waned.

Upon further examination of the sentiment score trends of the two models, it was observed that they follow similar patterns, with VADER reporting slightly higher numbers. The results indicated that a majority of the tweets had a neutral sentiment, followed by positive and negative sentiments in almost equal proportions. Interestingly, the neutral tweets seemed to have attracted more retweets and favorites, suggesting that users prefer to engage with tweets that are more objective and informative, rather than those that are overly emotional or controversial.

To gain a better understanding of the most popular vaccine among the users, a word cloud was generated from the tweets. The results showed that the Pfizer-BioNTech vaccine was the most commonly mentioned among the tweets. This could be attributed to its early availability and distribution in many countries, as well as its high efficacy rates. Overall, the insights gained from this analysis could be helpful in understanding public sentiment towards COVID-19 vaccination and could aid in improving future vaccination campaigns.

6. Data Privacy Approach

Performing sentiment analysis on Twitter data related to COVID-19 vaccination tweets may raise some concerns around data confidentiality, privacy, and GDPR compliance. Here are some approaches that can be used to address these issues:

- Anonymize data: Remove any personal identifiable information (PII) such as user names, profile pictures, and location data before conducting sentiment analysis.
- Use secure data transfer methods: Ensure that the data is transferred securely using encryption methods such as HTTPS or SSL.
- Obtain user consent: Obtain explicit consent from users before collecting and analyzing their tweets. This can be done by adding a notice on the data collection platform or by reaching out to users directly.
- Limit data access: Limit the number of individuals who have access to the data and ensure that those who have access follow strict confidentiality guidelines.
- Implement access controls: Implement access controls that prevent unauthorized access to the data.
- Use data pseudonymization: Replace identifiable information with pseudonyms to minimize the risk of data breaches.
- Conduct a Data Protection Impact Assessment (DPIA): DPIA is a risk assessment process that helps identify and mitigate any privacy or GDPR compliance risks associated with data processing activities.
- Follow GDPR regulations: Ensure that all data processing activities comply with GDPR regulations, including data subject rights, data retention, and data deletion requirements.

By following these approaches, that the sentiment analysis of COVID-19 vaccination tweets is conducted in a way that is compliant with data privacy and GDPR regulations while maintaining the confidentiality of the data.

7. Conclusion & Recommendations

In conclusion, the analysis of Twitter sentiment on COVID-19 vaccination tweets using VADER and TextBlob revealed that the majority of tweets had a neutral positive sentiment towards vaccination. Both sentiment analysis tools showed similar results with a high degree of agreement. However, it's important to note that this analysis has several limitations, including the fact that it was done on a limited dataset of tweets and that the accuracy of sentiment analysis tools may vary depending on the specific context and language used in the tweets.

Furthermore, it's important to consider data confidentiality and privacy issues when conducting sentiment analysis on Twitter data. Given that Twitter data is public, there may not be significant privacy concerns. However, GDPR regulations should be taken into consideration if personal data of users is being collected for the analysis.

Finally, this analysis was conducted locally and not connected to cloud resources. This may limit the scalability of the analysis, but it also reduces potential security concerns and ensures data privacy.

Overall, it is important to keep in mind that sentiment analysis is not a perfect science, and the results should always be interpreted with caution. Additionally, it may be useful to explore other types of analysis, such as topic modeling or network analysis, in order to gain a more comprehensive understanding of the dataset.

References

Akash (2021) *TextBlob | Making Natural Language Processing easy with TextBlob*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/10/making-natural-language-processing-easy-with-textblob/ (Accessed: 21 April 2023).

Alsaeedi, A. and Khan, M. (2019) 'A Study on Sentiment Analysis Techniques of Twitter Data', *International Journal of Advanced Computer Science and Applications*, 10, pp. 361–374. Available at: https://doi.org/10.14569/IJACSA.2019.0100248.

Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python*. O'Reilly Media, Inc. Available at: https://www.oreilly.com/library/view/natural-language-processing/9780596803346/ (Accessed: 21 April 2023).

Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2016) 'Annotate-Sample-Average (ASA): A New Distant Supervision Approach for Twitter Sentiment Analysis', in *22nd European Conference on Artificial Intelligence (ECAI)*. *22nd European Conference on Artificial Intelligence (ECAI)*, IOS Press, pp. 498–506. Available at: https://doi.org/10.3233/978-1-61499-672-9-498.

Hutto, C. and Gilbert, E. (2014) 'VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text', in *Proceedings of the International AAAI Conference on Web and Social Media*. *Eighth International AAAI Conference on Weblogs and Social Media*, pp. 216–225. Available at: https://doi.org/10.1609/icwsm.v8i1.14550.

Li, H., Dombrowski, L. and Brady, E. (2018) *Working Toward Empowering a Community: How Immigrant-Focused Nonprofit Organizations Use Twitter During Political Conflicts*. ACM Publications. Available at: https://doi.org/10.1145/3148330.3148336.

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. Cham: Springer International Publishing (Synthesis Lectures on Human Language Technologies). Available at: https://doi.org/10.1007/978-3-031-02145-9.

NLTK :: Natural Language Toolkit (no date). Available at: https://www.nltk.org/ (Accessed: 21 April 2023).

Paul, D. *et al.* (2017) 'Compass: Spatio Temporal Sentiment Analysis of US Election What Twitter Says!', in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery (KDD '17), pp. 1585–1594. Available at: https://doi.org/10.1145/3097983.3098053.

Praveen, S., Ittamalla, R. and Deepak, G. (2021) 'Analyzing the attitude of Indian citizens towards COVID-19 vaccine – A text analytics study', *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(2), pp. 595–599. Available at: https://doi.org/10.1016/j.dsx.2021.02.031.

Tang, D. *et al.* (2015) 'A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11), pp. 1750–1761. Available at: https://doi.org/10.1109/TASLP.2015.2449071.

Wang, H. *et al.* (2012) 'A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle', in *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, Korea: Association for

Computational Linguistics, pp. 115–120. Available at: https://aclanthology.org/P12-3020 (Accessed: 21 February 2023).