Digitus : Journal of Computer Science Applications Insight
Volume. 1 Issue 1 October 2023
Page No: 1-10

Sentiment Analysis of Public Opinion on the 2024 Presidential Election in Indonesia Using Twitter Data with the K-NN Method

Karno Diantoro[1], Ahmad Soderi[2], Abdur Rohman[3] , Anwar T. Sitorus[4]
Mercusuar College of Management and Informatics (STMIK Mercusuar), Indonesia
Coresspondent: karno@mercusuar.ac.id[1]

Received : August 18, 2023
Accepted : September 29, 2023
Published : October 13, 2023

Citation: Diantoro, K.., Soderi, A., Rohman, A., Sitorus, A.T. (2023). Sentiment Analysis of Public Opinion on the 2024 Presidential Election in Indonesia Using Twitter Data with the K-NN Method. Digitus : Journal of Computer Science Applications, 1(1), 1-10

ABSTRACT: Twitter is often used by the public as a platform to speak and express their opinions, especially in the context of the 2024 Presidential Election. Tweets related to the '2024 Presidential Election' can be used as a source of data for social media analysis to determine whether the expressed opinions tend to be positive or negative. The research process involves data collection of tweets, preprocessing, tokenization, class attribute determination, directory filling, sentiment analysis, and classification steps, including testing the value of k and testing the confusion matrix. The research and testing results show that the K-NN method successfully achieves a sentiment classification accuracy rate of 86.48%.

Keywords: K-NN, Election, Sentiment Analyst, Twitter

INTRODUCTION

Twitter is frequently used by the Indonesian population as a platform to express their opinions. Databoks reported that the number of Twitter users in Indonesia reached 18.45 million in January 2022, making Indonesia the 5th country in the world with the highest number of Twitter users. Twitter is one of the significant sources of data for tracking public opinions. It is one of the largest social media platforms globally, with millions of active users every day, resulting in a vast volume of data suitable for representative sentiment analysis. Most data uploaded on Twitter is public and can be accessed legally, without requiring special permission from users or violating their privacy. This data can be processed and used for social studies.

Indonesia is a country that operates a democratic system, reflected in the public's participation in government elections, both at the central and regional levels. General elections are held every five years. Presidential elections are always highly anticipated events and hot topics among the public, both in direct conversations and on social media. Diverse opinions emerge regarding every competing presidential candidate. The next is scheduled for 2024, although various predictions have already circulated regarding the candidate pairs who will participate.

The objective of this research is to examine the public's responses to the upcoming 2024 presidential election (pilpres). In social media analysis, tweets related to "Pilpres 2024" are used as a data source to determine whether public opinions tend to be positive or negative. The method

used in this analysis is the K-Nearest Neighbor. This research aims to leverage these opinions and understand the public's views on the 2024 presidential election (pilpres) in Indonesia based on opinions expressed by Twitter users.

METHOD

Literature Review

In 2021, Fajar Sodik Pamungkas and Iqbal Kharisudin conducted a research on Sentiment Analysis using SVM, NAIVE BAYES, and K-NN for Studying the Response of the Indonesian Society to the Covid-19 Pandemic on the Twitter Social Media platform. The Covid-19 pandemic has had significant impacts on various aspects of people's lives, necessitating physical distancing measures and changing societal norms. This has led to various opinions and responses from the public regarding the Covid-19 pandemic, which have been expressed on social media. To understand the sentiment of these public responses, sentiment analysis using machine learning algorithms is needed.

In this research, sentiment analysis of the Indonesian public's responses to the Covid-19 pandemic on the Twitter social media platform was conducted. The algorithms used included Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbor (K-NN). These three algorithms were then compared to determine which one was the most effective in classifying response data. Based on the average accuracy levels obtained using 10-Fold Cross Validation model evaluation, it was concluded that the SVM algorithm had a higher accuracy rate compared to Naive Bayes and K-NN. The average accuracy obtained was 90.01% for SVM with a linear kernel, 79.20% for Naive Bayes with a Laplace value of 1, and 62.10% for K-NN with a K value of 20 using an optimal kernele.

The procedure applied in this research involves several stages of the process, starting from collecting tweet data, data preprocessing, polarity calculation, sentiment classification, to model testing.

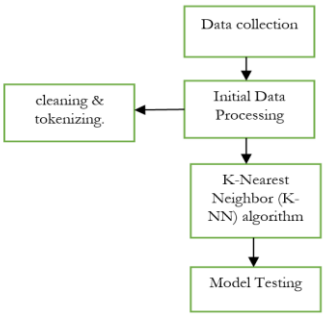

Figure 1. Research Flow

K-Nearest Neighbor (K-NN) is one of the popular algorithms. K-NN belongs to the instance-based learning group. The K-NN method is a lazy learning technique. The K-NN method has

several advantages that make it popular in machine learning, including: (1) Conceptual simplicity, where the core concept of K-NN is easy to understand; (2) K-NN only needs to store training data in memory for use in classification or prediction, as this algorithm is instance-based learning and does not require expensive training processes for parameter optimization; (3) The ability to handle complex data, where the K-NN method can be used to handle data with complex and nonlinear structures; (4) The ability to handle multiclass, allowing the use of K-NN for multiclass classification; (5) Because this algorithm does not assume a balanced class distribution, K-NN can make good predictions for minority classes.

Here are the weaknesses of the K-NN method to be noted: (1) Sensitivity to data scale, the K-NN method is highly sensitive to data scale because distance measurements between data depend on the units of measurement; (2) Influence of irrelevant attributes, K-NN does not automatically find the most relevant or important attributes in classification or prediction; (3) The K-NN method has high computational costs, especially for large datasets; (4) Issues with the number of neighbors (k). In the K-NN method, choosing the right number of neighbors (k value) is crucial; (5)K-NN does not explicitly model the relationships between attributes; this method only classifies or predicts similarity or distance between data. Consequently, K-NN may not be effective in identifying nonlinear relationships or complex interactions between attributes.

RESULT AND DISCUSSION

Model Testing

In the context of classification, measuring the accuracy of a classification model is a crucial aspect. Accuracy is used to assess how well the classification model can make accurate predictions. In this research, we conducted accuracy testing using cross-validation technique. In this technique, the dataset is divided into two main parts, namely the training set and the testing set. The training set is used to train the model, while the testing set is used to evaluate the model's performance.

The use of cross-validation technique with a certain number of iterations (epochs) is done to prevent overfitting and overlapping on the testing data. After that, the testing data is processed to create a confusion matrix. A Confusion Matrix is a matrix that contains the classification results performed by the system, both actual and predictive. For the case of binary classification, the results of the confusion matrix can be seen in the following table.

Table 1. Confusion Matrix

		Prediksi	
Aktual		Positif	Negatif
	Positif	*TP (True Positif)*	*FN (False Negatif)*
	Negatif	*FP (False Positif)*	*TN (True Negatif)*

From the analysis of values in the matrix (True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP)), we can determine accuracy, precision, and recall. Accuracy reflects how closely the classification aligns with the true values.
Accuracy is calculated by comparing the data that is correctly classified (TP and TN) with the total data.
Accuracy Formula = (TP + TN) / (TP + FP + FN + TN) x 100%
Description:

- TP: The number of true positive cases that are correctly classified as positive.
- FP: The number of negative cases incorrectly classified as positive.
- TN: The number of true negative cases that are correctly classified as negative.
- FN: The number of positive cases incorrectly classified as negative.

Precision is a measure of accuracy that indicates how closely the results of each iteration approximate the true values. Precision helps us evaluate how closely the answers given by the system match the requested information.

- Precision Positive Formula = TP / (TP + FP) x 100%
- Precision Negative Formula = TN / (TN + FN) x 100%

Recall, often referred to as sensitivity, is the percentage value that indicates how well a model can predict data into its actual classes.

- Recall Positive Formula = TP / (TP + FN) x 100%
- Recall Negative Formula = TN / (TN + FP) x 100%.

The application of the K-NN method in this research aims to classify sentiment towards the general election and determine the level of accuracy of sentiment analysis obtained from public comments on Twitter social media. The expected results include Accuracy, Positive Precision, Negative Precision, Positive Recall, and Negative Recall.

Crawling Data

The data collection process or Twitter data retrieval process uses the Twitter API Key. This Application Program Interface (API) is a service that consists of a set of commands, functions, components, and protocols provided by Twitter to facilitate the development of software systems. The number of Twitter data collected is 1000 tweets, and the successfully retrieved data covers the period from July 2, 2023, to July 4, 2023.

Preprocessing Data

The tweet data obtained from Twitter is still in its raw form; therefore, a data preprocessing process is needed to obtain clean and structured data for sentiment classification. The data preprocessing process consists of several stages, with three main stages as follows:

Cleaning Process

The steps involved in the data cleaning process include:

a. Attribute Selection: This stage involves the removal of irrelevant attributes in sentiment analysis.
b. Removal of Duplicate Data: In this stage, the initial data obtained from the previous Twitter data retrieval process (which amounted to 1000 data) will be examined to eliminate duplicate data, resulting in 881 remaining tweets to be used in the next stage.
c. Handling Missing Data: In this process, the obtained data will be checked for missing values or removed if they have no value. In this study, no missing values were found (Missing Value) in the tweet data.
d. Use of Subprocesses: The replacement process (Replace) is used as part of the subprocess in this stage. Subprocesses are used as containers for replacing specific values in the data.

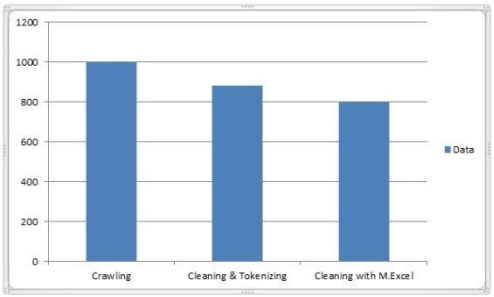

Figure 2. Data preprocessing results

Tokenization Process

In the tokenization process, each word in the text is segmented or separated into what is known as tokens. This process involves several steps, including: (1)Adding the Nominal to Text attribute, which serves as...; (2) Adding the Document from Data process, which functions as a container or storage for the tokenization process; (3) Adding the Wordlist process, which is used to calculate the frequency of data occurrences.

Classification with the K-NN Algorithm

In this stage, data that has undergone preprocessing and has become clean data will undergo classification using the K-NN algorithm. In this stage, the machine will be trained to recognize patterns in the existing data or documents, enabling it to classify data into three classes: positive, negative, and neutral.

Wordcloud Visualization

A wordcloud is a visual representation of text where the size of words in the text is depicted proportionally to their frequency. In a wordcloud, words that appear more frequently in the text will be displayed with a larger size, while words that appear less frequently will be displayed with a smaller size.

Figure 3. Wordcloud

Test Results

After the preprocessing stage is completed, the next step is to perform classification through the validation phase. The results of the testing using the K-Nearest Neighbor method will be evaluated. Below are the results of the confusion matrix generated by the K-Nearest Neighbor algorithm, which can be seen in Figure 4.

accuracy: 86.48% +/- 2.68% (micro average: 86.48%)

	true P	true N	class precision
pred. P	613	95	86.58%
pred. N	13	78	85.71%
class recall	97.92%	45.09%	

Figure 4. Test Results K-NN algorithm

The accuracy result of K-NN testing for sentiment analysis on Twitter social media is 86.48%. The precision for predicting the positive class is 86.58%, while for predicting the negative class is 85.71%. The recall obtained from true positives is 97.92%, while for true negatives, it is 45.09%.

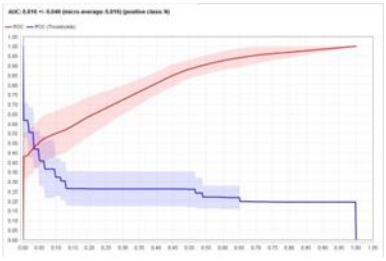

Figure 5. Graphic AUC

The Area Under the Curve (AUC) obtained is 0.816, indicating that the classification accuracy in this research falls into the category of "Good Classification. Analysis Results After conducting sentiment analysis using RapidMiner, it is evident that the initial data or training data underwent a reduction as various stages were carried out, starting from data collection (crawling) to classification using the K-NN algorithm. Here is a record of the data changes:

1) Data Collection (Crawling): The initial training data or raw data successfully collected using RapidMiner amounted to 1000 Twitter data.
2) Data Cleaning and Tokenization: After the cleaning process, the remaining data amounted to 881 tweets.
3) Further Data Cleaning Using Microsoft Excel: After removing hyperlinks and special characters that could not be eliminated by RapidMiner, there were 800 data left.

4) Through validation testing using the K-NN method with cross-validation, the average results were obtained for Positive Precision at 86.58%, Negative Precision at 85.71%, Positive Recall at 97.92%, and Negative Recall at 45.09%.

So, there were 800 test data used in the sentiment analysis stage, and the accuracy obtained with the K-NN method was 86.48%. This indicates a high level of positive sentiment toward the 2024 Presidential Election in Indonesia.

CONCLUSION

Based on the process stages explained in the previous chapter, conclusions can be drawn from the research results as follows:

1) The assistance of the Application Programming Interface (API) were used to easily retrieve tweet data from Twitter. The "search twitter" operator available in RapidMiner makes it easier for users to collect data, and obtaining Twitter API can be done without complications.
2) This research aimed to evaluate the positive and negative sentiments on Twitter related to the 2024 presidential election in Indonesia. To represent this, text mining and text classification using the K-NN method were conducted to classify sentiment labels from the dataset. The test results show that the algorithm used in the study, K-Nearest Neighbor (K-NN), performed well. This is evidenced by the accuracy rate reaching more than 50%. Accuracy measures the extent to which the obtained values match the true values by the K-NN algorithm, which is 86.48%.

3) Based on the visualization of the word cloud by selecting 30 frequently mentioned or discussed words by the public, the words "pemilu" (elections) and "anies" stood out. Both of these words had the highest occurrence rates, with "pemilu" appearing 799 times and "anies" appearing 415 times, each with a 100% occurrence rate compared to other words.

4) In the validation test of the K-NN method using cross-validation, an average value of 86.58% was found for Positive Precision, which measures the suitability of the data portion taken with the required positive information. Negative Precision was 85.71%, measuring the suitability of the data portion taken with the required negative information. Positive Recall was 97.92%, measuring the system's success rate in rediscovering positive information. Negative Recall was 45.09%, measuring the system's success rate in rediscovering negative information.

REFERENCE

Alsabban, M. (2021). Comparing two sentiment analysis approaches by understand the hesitancy to COVID-19 vaccine based on Twitter data in two cultures. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3462741.3466671

Aquino, P. A., López, V. F., Moreno, M. N., Muñoz, M. D., & Rodríguez, S. (2020). Opinion Mining System for Twitter Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12344 LNAI*. https://doi.org/10.1007/978-3-030-61705-9_38

Azzouza, N., Akli-Astouati, K., & Ibrahim, R. (2020). Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations. *Advances in Intelligent Systems and Computing*, *1073*. https://doi.org/10.1007/978-3-030-33582-3_41

Bernal, C., Bernal, M., Noguera, A., Ponce, H., & Avalos-Gauna, E. (2021). Sentiment Analysis on Twitter About COVID-19 Vaccination in Mexico. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13068 LNAI*. https://doi.org/10.1007/978-3-030-89820-5_8

Brito, K., Filho, R. L. C. S., & Adeodato, P. (2022). Please stop trying to predict elections only with Twitter. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3543434.3543648

Choudhary, N., Singh, R., Bindlish, I., & Shrivastava, M. (2023). Sentiment Analysis of Code-Mixed Languages Leveraging Resource Rich Languages. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13397 LNCS*. https://doi.org/10.1007/978-3-031-23804-8_9

Dutta, R. (2021). To Find the Best-Suited Model for Sentiment Analysis of Real-Time Twitter Data. *Advances in Intelligent Systems and Computing*, *1165*. https://doi.org/10.1007/978-981-15-5113-0_34

Feitosa, M. F., Rocha, S., Gonçalves, G. D., Ferreira, C. H., & Almeida, J. M. (2022). Sentiment Analysis on Twitter Repercussion of Police Operations. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3539637.3558050

Gautam, J., Atrey, M., Malsa, N., Balyan, A., Shaw, R. N., & Ghosh, A. (2021). Twitter Data Sentiment Analysis Using Naive Bayes Classifier and Generation of Heat Map for Analyzing Intensity Geographically. In *Advances in Intelligent Systems and Computing* (Vol. 1319). https://doi.org/10.1007/978-981-33-6919-1_10

Gupta, I., & Joshi, N. (2022). A Review on Negation Role in Twitter Sentiment Analysis. In *International Journal of Healthcare Information Systems and Informatics* (Vol. 16, Issue 4). https://doi.org/10.4018/IJHISI.20211001.oa14

Ilyas, S. H. W., Soomro, Z. T., Anwar, A., Shahzad, H., & Yaqub, U. (2020). Analyzing brexit's impact using sentiment analysis and topic modeling on twitter discussion. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3396956.3396973

Jiang, T., Wang, J., Liu, Z., & Ling, Y. (2020). Fusion-Extraction Network for Multimodal Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12085 LNAI*. https://doi.org/10.1007/978-3-030-47436-2_59

Joshi, D. J., Kankurti, T., Padalkar, A., Deshmukh, R., Kadam, S., & Vartak, T. (2021). Performance Analysis of Different Models for Twitter Sentiment. *Advances in Intelligent Systems and Computing*, *1311 AISC*. https://doi.org/10.1007/978-981-33-4859-2_11

Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12844 LNCS*. https://doi.org/10.1007/978-3-030-84060-0_11

Kaur, C., & Sharma, A. (2021). COVID-19 Sentimental Analysis Using Machine Learning Techniques. *Advances in Intelligent Systems and Computing*, *1299 AISC*. https://doi.org/10.1007/978-981-33-4299-6_13

Kumar, P., Reji, R. E., & Singh, V. (2022). Extracting Emotion Quotient of Viral Information Over Twitter. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13145 LNCS*. https://doi.org/10.1007/978-3-030-94876-4_15

Li, Q., Zhang, J., Guo, J., Li, J., & Kang, C. (2021). Evaluating Performance of NBA Players with Sentiment Analysis on Twitter Messages. *ACM International Conference Proceeding Series*.

https://doi.org/10.1145/3501774.3501796
Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. *Artificial Intelligence Review*, *54*(7). https://doi.org/10.1007/s10462-021-09973-3
Limboi, S., & Dioşan, L. (2020). Hybrid Features for Twitter Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12416 LNAI*. https://doi.org/10.1007/978-3-030-61534-5_19
Liu, H., & Tan, E. (2022). Tweet Sentiment Extraction Using Byte Level Pretrained Language Modelĝ. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3529836.3529941
Lohar, P., Xie, G., Bendechache, M., Brennan, R., Celeste, E., Trestian, R., & Tal, I. (2021). Irish Attitudes Toward COVID Tracker App & Privacy: Sentiment Analysis on Twitter and Survey Data. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3465481.3469193
Murakami, H., Ejima, N., & Kumagai, N. (2020). Self-understanding support tool using twitter sentiment analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12144 LNAI*. https://doi.org/10.1007/978-3-030-55789-8_29
Nandy, H., & Sridhar, R. (2021). Filtering-Based Text Sentiment Analysis for Twitter Dataset. *Advances in Intelligent Systems and Computing*, *1133*. https://doi.org/10.1007/978-981-15-3514-7_77
Nouira, A. Y., Bouchakwa, M., & Jamoussi, Y. (2023). Bitcoin Price Prediction Considering Sentiment Analysis on Twitter and Google News. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3589462.3589494
Nugroho, K. S., Sukmadewa, A. Y., Dw, H. W., Bachtiar, F. A., & Yudistira, N. (2021). BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3479645.3479679
Obaidi, M., & Klünder, J. (2021). Development and application of sentiment analysis tools in software engineering: A systematic literature review. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3463274.3463328
Ricci, R. D., Faria, E. R., Miani, R. S., & Gabriel, P. H. R. (2021). Social Security Reform in Brazil: A Twitter Sentiment Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12926 LNCS*. https://doi.org/10.1007/978-3-030-86611-2_11
Rocha, R. S., Saraiva, L. A., Castro, A. F. De, & Silva, P. D. A. (2020). Sentiment analysis of Twitter data about blockchain technology. *ACM International Conference Proceeding Series*, *Part F166737*. https://doi.org/10.1145/3401895.3401913
Selmene, S., & Kodia, Z. (2020). Recommender System Based on User's Tweets Sentiment Analysis. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3409929.3414744
Silva, J., Cera, J. M., Vargas, J., & Lezama, O. B. P. (2021). Sentiment analysis in twitter: Impact of morphological characteristics. *Advances in Intelligent Systems and Computing*, *1237 AISC*. https://doi.org/10.1007/978-3-030-53036-5_29
Vaseeharan, T., & Aponso, A. (2020). Review on Sentiment Analysis of Twitter Posts about News Headlines Using Machine Learning Approaches and Naïve Bayes Classifier. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3384613.3384650
Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment Analysis of Twitter Data. In *Applied Sciences (Switzerland)* (Vol. 12, Issue 22). https://doi.org/10.3390/app122211775
Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7).

https://doi.org/10.1007/s10462-022-10144-1

Xie, Y., Wang, T., Zhang, H., & Yan, T. (2022). Analyzing the Rate of Increase in Vaccines Administrated Versus Twitter Sentiment Analysis. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3543106.3543119

Yadlapalli, S. S., Reddy, R. R., & Sasikala, T. (2020). Advanced Twitter Sentiment Analysis Using Supervised Techniques and Minimalistic Features. *Advances in Intelligent Systems and Computing*, *1097*. https://doi.org/10.1007/978-981-15-1518-7_8

Yang, L., Yu, J., Zhang, C., & Na, J. C. (2021). Fine-Grained Sentiment Analysis of Political Tweets with Entity-Aware Multimodal Network. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12645 LNCS*. https://doi.org/10.1007/978-3-030-71292-1_31

Zhu, W., & Hu, T. (2021). Twitter Sentiment analysis of covid vaccines. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3480433.3480442