Sentiment Analyst on Twitter Using the K-Nearest Neighbors (KNN) Algorithm Against Covid-19 Vaccination

Suprayogi*[1], **Christy Atika Sari**[2]
Dian Nuswantoro University, Imam Bonjol 207 Semarang 50131
E-mail : suprayogi@dsn.dinus.ac.id[1], christy.atika.sari@dsn.dinus.ac.id[2]*
**Corresponding author*

Eko Hari Rachmawanto[3]
Dian Nuswantoro University, Imam Bonjol 207 Semarang 50131
E-mail : eko.hari@dsn.dinus.ac.id[3]

Abstract - The corona virus (2019-nCoV), commonly known as COVID-19 has been officially designated as a global pandemic by the WHO. Twitter, is one of the social media used by many people and is popular among internet users in expressing opinions. One of the problems related to Covid-19 and causing a stir is the procurement of the Covid-19 vaccine. The procurement of the vaccine caused various opinions in Indonesian society, where the uproar was also quite busy being discussed on Twitter and even became a Trending Topic. The opinions that appear on Twitter will then be used as data for the Sentiment Analysis process. One of the members of the House of Representatives (DPR), namely Ribka Tjiptaning was also included in the Trending Topic list on Twitter for refusing to receive the Covid-19 vaccine. Sentiment analysis itself is a computational study of opinions, sentiments and emotions expressed textually. Sentiment analysis is also a technique to extract information in the form of a person's attitude towards an issue or event by classifying the polarity of a text. Research related to Sentiment Analysis will be examined by dividing public opinion on Twitter social media into positive and negative sentiments, and using the K-Nearest Neighbor (KNN) algorithm to classify public opinion about COVID-19 vaccination. In the testing section, the Confusion Matrix method is used which then results in an accuracy of 85%, precision of 100%, and recall of 78.94%.

Keywords –Sentiment Analysis, Covid-19, K-Nearest Neighbor, Confusion Matrix

1. INTRODUCTION

The outbreak of a new disease caused by the corona virus (2019-nCoV) or commonly known as COVID-19 was officially declared a global pandemic by the World Health Organization (WHO) on March 11, 2020. Although the epicenter of the virus outbreak at the end of 2019 was from the Chinese state of Wuhan, the virus has now spread globally, with more than 41.5 million cases and more than 1.1 million deaths annually. In Indonesia itself, President Joko Widodo announced the first COVID-19 case that entered Indonesia on March 2, 2020. According to Liu et al, Covid-19 vaccination is one of the effective ways to deal with the spread of the Corona virus. Because after receiving the Covid-19 vaccination, the immune body of the human body will be immune and will get the benefits of being protected from Covid-19[1], [2], but on the other hand, vaccination is also in protecting others around us so that it can reduce the expansion and spread of the Covid-19 virus. Vaccination campaign plans must consider all aspects, from the feasibility of using the vaccine, the risks after use, to the various stages and procedures from vaccination to outreach to the public[3]–[6].

One of the social networking media that is often used by many people and is very popular among internet users in giving their opinions is Twitter. Indonesia is one of the countries that has quite a lot of daily active Twitter users. Based on data from Hootsuite, Indonesia is in 8th place with a reach of 10 million users[7]–[9]. The procurement of the Covid-19 vaccine has caused mixed opinions in Indonesian society. On Twitter social media, the corona vaccine had become a trending topic because it was widely discussed by the Indonesian people. Opinions that are on Twitter will then become data for sentiment analysis. One of the members of the House of Representatives (DPR) Ribka Tjiptaning who became a trending topic on Twitter refused to receive the Covid-19 vaccine, even though he is 63 years old and prefers to pay a fine, he was given a sanction by the government of 5 million rupiah on the grounds that Bio Farma had not tested of the Covid-19 vaccine, highlighting the incidence of the polio vaccine and elephantiasis vaccine in Indonesia. Therefore, it can be assumed that if the Covid-19 vaccine is to be used in humans, further evidence is needed. Sentiment analysis was applied in this study to analyze the opinions, feelings, views, and preferences of individuals regarding the COVID19 vaccination event by collecting data from Twitter users who discussed the topic of vaccination against COVID19[10]–[12]. The spread of this pandemic throughout the world and the imposition of restrictions on social interaction affect people's social conditions, the circulation of problems, both positive and negative, and even creates a big panic in social networks[7], [13].

The K-Nearest Neighbor algorithm[14], [15] is a classification algorithm because it is easy to implement, using data that has labels so that during the grouping process it becomes easier to get into the most appropriate class. Has ease in the aspect of translating the results, accuracy and calculation time of predictions. The K-NN algorithm has several advantages including being proven in accordance with the calculations applied in an application and achieving good accuracy results [6], [14], [16], the K-NN algorithm has a pretty good performance as a classification shown by several researchers who use it, especially classification in text. This is evidenced through a study entitled Twitter Sentiment Analysis Against the 2019 Indonesian Presidential Candidates using the K-NN Method which obtained an accuracy rate of the K-NN method reaching 83.33%.

2. RESEARCH METHOD

2.1. K-Nearest Neighbor (KNN) Algorithm

The K-Nearest Neighbors algorithm is an approach to finding cases by calculating the proximity between new cases and old cases based on the weight matching of a number of existing features as pattern recognition that is commonly used for classification and regression purposes[17], [18]. The classification process carried out in text mining aims to classify data into several groups. The process of grouping data refers to the data that has been known in advance the group or class. Data that does not have a group is determined by a comparison process with data that is already known to the group. K-Nearest Neighbor (K-NN) is a classification technique that makes accurate predictions on test data based on the comparison of K nearest neighbors. Parameter K in K-Nearest Neigbor, K is the number of closest neighbors involved and has an influence in determining the prediction results.

K-Nearest Neighbor is an instance-based classification method that selects a training object with the nearest neighbor attribute. The nature of this neighbor is obtained from the calculation of the value of similarity or dissimilarity. KNN uses a method to calculate dissimilarity values (Euclidian, Manhattan, Square Euclidian, dil). Near or far neighbors are usually calculated based on the eludian distance. KNN will choose the K closest neighbors by looking at the number of class occurrences in the selected K neighbors to determine the classification results. The best value of k for this algorithm depends on the data, usually a high value of k will reduce the effect of noise on the application. Good & values can be selected by parameter optimization, for example by cross-validation. The class that appears the most will be the class resulting from the classification. The best K value for this algorithm depends on the data in general, a high K value will reduce the effect of noise on the classification. If an unlabeled object is given, the algorithm searches for the same or neighboring objects in the search space, and assigns a label to the unlabeled object based on the nearest neighbor attribute. The same concept can also be applied to sequences of observations, such as measuring current levels. The K-NN algorithm identifies K past data sequences that are most similar to the pattern being examined. The closest combination of values makes an expected prediction of the expected future value based on a time step. The general closeness is between values 0 and 1. A value of 0 means that the two cases are not at all similar, and the value of 1 case is almost exactly similar. The process of calculating the distance between the two cases is carried out using the following equation.

$$similarity\ (T, S) = \frac{\sum_{i=1}^{n}(T_i, S_i) * W_i}{W_i} \qquad (1)$$

Where:

T = New case
S = Cases that are in deviation
n = Number of attributes in each case
i = Individual attributes between 1 to n
f = The similarity function of attribute i between cases T and S
w = The weight assigned to the attribute i

2.2. *Confussion Matrix*

The performance of the classification system describes how well the data classification system is. Confusion matrix is a method that can be used to measure the performance of a classification method. Basically, the confusion matrix contains information that compares the results of the classification performed by the system with the classification results that should be obtained. The confusion matrix can be used to evaluate the algorithm performance of Machine Learning (ML)[3]. The Confusion Matrix represents the predictions and actual (actual) conditions of the data generated by the Machine Learning (ML) algorithm. Based on the Confusion Matrix, it can determine Accuracy, Precision, Recall Specificity and F1 score. When measuring the performance of an algorithm using a confusion matrix, there are 4 terms to represent the results of the classification process. The four terms are TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). The TN value is the number of negative data that is correctly detected, while the FP is negative data but detected as positive data. Meanwhile, TP is positive data that is detected correctly. FN is the opposite of True Positive, so the data is positive, but is detected as negative data.

2.3. *RapidMiner*

Rapidminer is an open source software which is one of the solutions for analyzing predictive analysis, text mining, and data mining. Rapidminer uses various descriptive and predictive techniques in providing knowledge to users so that users can make the best decisions. Rapidminer is a stand-alone software and has functions for data analysis and can be used as a data mining machine that can be integrated into its own products. Rapidminer is written using the Java programming language so that it can work on all operating systems. Rapidminer has the following properties:

a. Developed using the Java programming language so that it can run on various operating systems.
b. The knowledge discovery process is modeled as operator trees.
c. XML representation, internal to ensure the standard format of data exchange.
d. The scripting language allows for large-scale experiments and automation of experiments.
e. Multi-layer concept to ensure efficient data display and guarantee data handling.
f. It has a GUI, command line mode, and java API that can be called from other programs.

2.4. Text Preprocessing

Text Preprocessing is a process that functions to clean text or words before further processing is carried out. Unstructured data and still contains noise such as punctuation marks, affixes, numbers, special characters, and others[19], [20]. At the text preprocessing stage so that the data used can be ready to be processed in the next phase.

2.5. TF-IDF

The TF-IDF (term frequency-inverse document frequency) stage plays a role in determining the terms or keywords that characterize a document that can distinguish documents from one another in one corpus[21], [22]. TF-IDF works by increasing in proportion to the number of times a word appears in the document, but offset by the number of documents containing the same keyword. In text mining, feature selection is the most important stage that has a very significant role in the accuracy of text analytics, because feature selection is a process used to remove or delete irrelevant features contained in a dataset. There are four most widely used approaches in the feature selection process, namely Document Frequency (DF), Term Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency/ Inverse Document Frequency (TFADF).

3. RESULTS AND DISCUSSION

3.1. Data Preparation

The data used is data taken from Twitter by utilizing Rapidminer software. As many as 300 data were taken which were divided into two equal numbers, namely 150 data were positive opinions, and the remaining 150 data were negative opinions.

Table 1. The Division of Positive and Negative Words

Positive Words (English)	Negative Words (English)
Aman (Safe)	Kecewa (Disappointed)
Efektif (Effective)	Efeksamping (Side Effects)
Siap (Ready)	Hoax (Hoax)
Mandiri (Independent)	Tergesa-gesa (Hurry)
Gratis (Free)	Takut (Afraid)
Terbaik (Best)	Mati (Dead)
Ampuh (Powerful)	Tidak perlu (No Need)
Percaya (Belive)	Menolak (Reject)
Terjangkau (Affordable)	Terburu-buru (In a hurry)
Tersedia (Availabla)	Meninggal (Die)

Menjaga (Guard)	Konspirasi (Conspiracy)
Mendukung (Support)	Meragukan (Doubt)
Halal (Halal)	Kecemasan (Worry)
Maju (Up)	Bingung (Confused)

In Table 1, several sample words that will be used as markers of a sentence are shown are positive or negative sentences. In the positive word category, if the word is not in one sentence, it is not included in the positive class, as well as the rules that apply to negative words.

Then, the data that has been obtained is applied to the pre-processing process with the following steps:

a. Data Cleansing, which is the process of removing unused symbols such as: # (hashtag), @ (at), RT (ReTweet), tags, links, and other scripts.
b. Case Foalding, which is the process of changing all uppercase letters to lowercase and the process of separating words.
c. Tokenization, which is the process of changing sentences into tokens (separation of words that make words into original words).
d. Stop-word Removal, which is the process of removing unimportant words.
e. Steaming, which is the process of changing words in terms into basic forms by removing affixes.

Table 2. Preprocessing Data Cleansing and Case Folding

Doc	Data Cleansing	Case Folding	Class
D1	AdamPrabata KABAR BAIK Booster vaksinModernaterbuktimampumeningkatkanantibodi	adamprabatakabar baik booster vaksinmodernaterbuktimampumeningkatkanantibodi	Positive
D2	Alhamdulillah vaksinkeduaaman	alhamdulillahvaksinkeduaaman	Positive
D3	Buat apavaksinkalau tidak ngefek	buat apavaksinkalau tidak ngefek	Negative
D4	Habisvaksintapimalahdemamberhari-hari	habisvaksintapimalahdemamberhari-hari	Negative
D5	Ayo lawancovid dengan vaksinkeduasinovacwalaupundemam	ayolawancovid dengan vaksinkeduasinovacwalaupundemam	????

After the document or data has passed the data cleansing and case folding stages, the next process is tokenization, stop-word removal and stemming. So get the results as in table 3.

Table 3. Preprocessing Data Tokenization to Stemming

Doc	Tokenization	Stop-word Removal	Stemming	Class
D1	“adamprabata”, “kabar”, “baik”, “booster”, “vaksin”, “moderna”, “terbukti”, “mampu”, “meningkatkan”, “antibody”	“adamprabata”, “kabar”, “baik”, “booster”, “vaksin”, “moderna”, “bukti”, “mampu”, “tingkat”, “antibodi”	“adamprabata”, “kabar”, “baik”, “booster”, “vaksin”, “moderna”, “bukti”, “mampu”, “tingkat”, “antibodi”	Positive
D2	“alhamdulillah”, “vaksin”, “kedua”, “aman”	“alhamdulillah”, “vaksin”, “dua”, “aman”	“alhamdulillah”, “vaksin”, “dua”, “aman”	Positive
D3	“buat”, “apa”, “vaksin”, “kalo”, “tidak”, “ngefek”	“buat”, “apa”, “vaksin”, “kalau”, “tidak”, “ngefek”	“buat”, “apa”, “vaksin”, “kalau”, “tidak”, “ngefek”	Negative
D4	“habis”, “vaksin”, “tapi”, “malah”, “demam”, “berhari”, “hari”	“habis”, “vaksin”, “tapi”, “malah”, “demam”, “hari”	“habis”, “vaksin”, “tapi”, “malah”, “demam”, “hari”	Negative
D5	“ayo”, “lawan”, “covid”, “vaksin”, “kedua”, “sinovac”, “walaupun”, “demam”	“ayo”, “lawan”, “covid”, “vaksin”, “dua”, “sinovac”, “walaupun”, “demam”	“ayo”, “lawan”, “covid”. “vaksin”, “dua”, “sinovac”, “walaupun”, “demam”	????

After the preprocessing process is complete, the next step is to calculate the TF and IDF in each term or word that represents each document. The frequency of a word in a particular document indicates its importance in the document. The frequency of documents containing the word indicates how common the word is. In this way, if a word appears frequently in the document, and the entire document containing that word appears infrequently in the document set, the weight of the word-document relationship will be high. And to get the IDF results, the calculation process uses the following formula:

$$IDF(w) = log\left(\frac{N}{DF}\right) \qquad (2)$$

After getting the IDF value, the TF-IDF value will be searched again, namely the multiplication between the results of the frequency of occurrence of words in each document (TF) with the weighting of words in all documents (IDF). The results of the multiplication of TF and IDF are in table 4.

Table 4. TF-IDF . Multiplication Results

Wdf=TF*IDF						
No	TERM	D1	D2	D3	D4	D5
1	adamprabata	0.69897	0	0	0	0
2	kabar	0.69897	0	0	0	0
3	baik	0.69897	0	0	0	0
4	booster	0.69897	0	0	0	0
5	vaksin	0	0	0	0	0
6	moderna	0.39794	0	0	0	0.39794
7	bukti	0.69897	0	0	0	0
8	mampu	0.69897	0	0	0	0
9	tingkat	0.69897	0	0	0	0
10	antibodi	0.69897	0	0	0	0
...	...	...	...	...	...	...
27	sinovac	0	0	0	0	0.69897
28	demam	0	0	0	0	0.69897

After getting the IDF value, the TF-IDF value will be searched again, namely the multiplication between the results of the frequency of occurrence of words in each document (TF) with the weighting of words in all documents (IDF). After getting the results of the document vector similarity, the next process is to calculate the length of the vector by squared the result of the document vector similarity for each term then add up the square value of each document and then calculate the root of that number as shown in Table 5.

Table 5. Results of Document Vector Similarities

No	D1	D2	D3	D4
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0

5	0	0	0	0
6	0.158356	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
…	…	…	…	…
27	0	0	0	0
28	0	0	0	0
Total	0.158356	0.158356	0	0.158356

After getting the results of the vector length, the next step is to calculate cosine similarity, to compare the similarities between documents. For the process of calculating cosine similarity by doing scalar multiplication between queries. After getting the results of the vector length, the next step is to calculate cosine similarity, to compare the similarities between documents. The process of calculating cosine similarity by doing scalar multiplication (D1, D2, D3, Dn) between queries (Dx).

Table 6. Document Vector Similarity Results

No	D1	D2	D3	D4	D5
1	0.488559	0	0	0	0
2	0.488559	0	0	0	0
3	0.488559	0	0	0	0
4	0.488559	0	0	0	0
5	0	0	0	0	0
6	0.158356	0	0	0	0.158356
7	0.488559	0	0	0	0
8	0.488559	0	0	0	0
9	0.488559	0	0	0	0
10	0.488559	0	0	0	0
…	…	…	…	…	…
27	0	0	0	0	0.488559
28	0	0	0	0	0.488559
Total	4.066829	1.135474	2.442795	2.112593	2.429305
The Final Result	2.016638	1.065586	1.562944	1.453476	1.558623

Continuing the process of calculating Cosine Similarity as follows:

$$\mathrm{CosSin}(\mathrm{D5},\mathrm{D1}) = \frac{0{,}158356}{1.558623 \times 2.016638} = 0{,}050381$$

$$\mathrm{CosSin}(\mathrm{D5},\mathrm{D2}) = \frac{0{,}158356}{1.558623 \times 1.065586} = 0{,}095347$$

$$\mathrm{CosSin(D5, D3)} = \frac{0}{1.558623 \times 1.562944} = 0$$

$$\mathrm{CosSin(D5, D4)} = \frac{0{,}158356}{1.558623 \times 1.453476} = 0{,}069901$$

The results obtained from cosine similarity will be sorted from the largest value to the smallest value. Next, a class will be given according to the class labeling at the beginning. Here is the table of cosine similarity results:

Table 7. Cosine Similarity Results

Doc	Sim(x, dn)	Class
D1	0.050381	POSITIVE
D2	0.095347	POSITIVE
D3	0	NEGATIVE
D4	0.069901	NEGATIVE

In calculating the classification using the K-NN method, the first thing to do is determine the value of K. Next, take the similarity results from the K value starting from the highest value.

Table 8. Order of Largest to Smallest Values

1	2	3	4
D2	D4	D1	D3
0.095347	0.069901	0.050381	0

3.2. KNN Algorithm

In calculating the classification using the K-NN method, the first thing to do is determine the value of K. Then take the similarity results from the K value starting from the highest value. After getting the results from the similarity of a specified number of K values, you can determine whether the classification results are included in the Positive or Negative class.

$$P(x, c_m) = \sum\nolimits_{i=1}^{m} SIM(X, d_j) \in c_m \qquad (3)$$

Probability against popular class:
P(x, Positif) = 0.050381+ 0.095347= 0,14893 (4)

Probability of unpopular class:
P(x, Negatif) = 0,097097 + 0 = 0,097097 (5)

Based on the test data "let's fight covid with the second Sinovac vaccine even though it has a fever" is a positive sentiment because from the highest k = 3, the probability of D5 against the Positive class is greater than that of the Negative class, so the sentiment on D5 is included in the Positive class category.

3.3. Confusion Matrix Test

Accuracy testing to determine the level of accuracy is carried out by the system using the K-Nearest Neighbor algorithm to classify public opinion about COVID-19 vaccination. To find out the value of the accuracy of the testing data used as much as 87 and training data as much as 224.

Table 9. Confusion Matrix Test Results

No	Sentiment	Sentiment Prediction	Positive Confidence	Negative Confidence	Text
1	Positive	Positive	1.0	0.0	Beruntunglah Yang Belum Vaksin
2	Negative	Positive	1.0	0.0	RT @Fatihah43840153: @LtdAkbarTemankuhabisvaksin sekarang kritis, kenaautoimun...
3	Positive	Positive	1.0	0.0	RT @DPP_PPP: PPP FasilitasiTes Antigen, Vaksin Gratis dandirikanposko untuk PesertaMuktamar NU
4	Negative	Positive	1.0	0.0	RT @S4fitr1HS: PemerintahanJokowiLawanPandemi @KemenkeuRI Sri Mulyanimengatakaninsentiffiskalbeadancukai untuk sektorkesehatan s…
5	Negative	Positive	1.0	0.0	RT @BuKasunNdeso: Lo.. kok bisa kelenjar di leherbengkak. Bukannyavaksinaman..?? Eh kalongantukan atau tiba2 seringtertidurapajuga…
6	Positive	Positive	1.0	0.0	keberhasilanvaksin
7	Positive	Positive	1.0	0.0	@nct_menfess GA KUAT MAJU LO LI ZENO GUA UDAH VAKSIN 2 KALI
8	Positive	Positive	1.0	0.0	Jog, info vaksindosis ke-2 astra area bantul/jogjahari ini dong. Urgent, syarat naikkereta. Padahalsebelumevaksinpertamatok gpp :(
9	Positive	Positive	1.0	0.0	@milkkit_ta Di aku bisa kacetaksertifikatvaksin. 10k udhbolakbalik. Pengirimandrjaksel. Data di jaminaman. Bisa via shopeejglg gratis ongkirxixi
10	Positive	Positive	1.0	0.0	SahalSabililkalaudibaca summary a.ikelnyasyberkesimpulan perlu development vaksin baru untuk omicron ini.. benerbegtuyapak?
…	…	…	…	…	…
86	Negative	Negative	0.0	1.0	Sungguhbiadab jika rakyatdijadikan korban bisnisvaksin.
87	Positive	Positive	1.0	0.0	tidak anti vaksintapi anti pemaksaankarena sekarang bukan zaman Romusa. Terima kasih

Table 10. Confusion Matrix Testing

	True Positif	True Negatif
False Positif	55	12
False Negatif	0	26

- Accuracy

$$= \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$= \frac{45 + 26}{45 + 26 + 0 + 12} \times 100\%$$

$$= \frac{71}{83} \times 100\%$$

$$= 85\%$$

- Precision

$$= \frac{TP}{TP + FP} \times 100\%$$

$$= \frac{45}{45 + 0} \times 100\%$$

$$= \frac{45}{45} \times 100\%$$

$$= 100\%$$

- Recall

$$= \frac{TP}{TP + FN} \times 100\%$$

$$= \frac{45}{45 + 12} \times 100\%$$

$$= \frac{45}{57} \times 100\%$$

$$= 78{,}94\%$$

It has been tested using the Confusion Matrix method with 87 testing data and 224 training data, getting results of 85% accuracy, 100% precision and 78.94% recall.

4. CONCLUSION

Accuracy testing was carried out using the confusion matrix method to obtain an accuracy of 85%, precision of 100% and recall of 78.94% from testing data of 83 data and 224 training data. The K-Nearest Neighbors algorithm used can capture public opinion related COVID-19 vaccines can be captured on Twitter social media, such as public discussions about vaccines, halal certification of vaccines, proper use of vaccines, vaccine prices, and to general public talks such as functions & objects of vaccination.

REFERENCES

[1] D. Ukkaz, "SENTIMENT ANALYSIS OF COVID-19 VACCINE WITH DEEP LEARNING," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 12, pp. 4513–4521, 2022.

[2] N. M. Abdulkareem, A. Mohsin Abdulazeez, D. Qader Zeebaree, and D. A. Hasan, "COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms," *Qubahan Acad. J.*, vol. 1, no. 2, pp. 100–105, 2021.

[3] A. Winanto and C. Budihartanti, "Comparison of the Accuracy of Sentiment Analysis on the Twitter of the DKI Jakarta Provincial Government during the COVID-19 Vaccine Time," *J. Comput. Sci. an Eng.*, vol. 3, no. 1, pp. 14–27, 2022.

[4] N. A. Azeez, O. E. Victor, and U. E. Junior, "SENTIMENT ANALYSIS OF COVID-19 TWEETS," *FUDMA J. Sci.*, vol. 5, no. 1996, p. 6, 2021.

[5] R. K. BANIA, "Heterogeneous Ensemble Learning Framework for Sentiment Analysis on COVID-19 Tweets," *INFOCOMP*, vol. 20, no. 02, 2021.

[6] F. M. J. M. Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, pp. 463–470, 2021.

[7] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012045, 2021.

[8] N. G. Ramadhan and F. D. Adhinata, "Sentiment analysis on vaccine COVID-19 using word count and Gaussian Naïve Bayes," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, p. 1765, 2022.

[9] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer," *3rd 2021 East Indones. Conf. Comput. Inf. Technol. EIConCIT 2021*, pp. 122–127, 2021.

[10] N. S. Sattar and S. Arifuzzaman, "Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the usa," *Appl. Sci.*, vol. 11, no. 13, 2021.

[11] S. Nyawa, D. Tchuente, and S. Fosso-Wamba, "COVID-19 vaccine hesitancy: a social media analysis using deep learning," *Ann. Oper. Res.*, 2022.

[12] A. M. Almars, E. S. Atlam, T. H. Noor, G. ELmarhomy, R. Alagamy, and I. Gad, "Users opinion and emotion understanding in social media regarding COVID-19 vaccine," *Computing*, vol. 104, no. 6, pp. 1481–1496, 2022.

[13] A. Umair and E. Masciari, "Sentimental and spatial analysis of COVID-19 vaccines tweets," *J. Intell. Inf. Syst.*, 2022.

[14] S. Hota and S. Pathak, "KNN classifier based approach for multi-class sentiment analysis of twitter data," *Int. J. Eng. Technol.*, vol. 7, no. 3, p. 1372, Jul. 2018.

[15] T. Mustaqim, K. Umam, and M. A. Muslim, "Twitter text mining for sentiment analysis on

government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm," *J. Phys. Conf. Ser.*, vol. 1567, no. 3, pp. 8–15, 2020.

[16] S. Kaur, G. Sikka, and L. K. Awasthi, "Sentiment Analysis Approach Based on N-gram and KNN Classifier," *ICSCCC 2018 - 1st Int. Conf. Secur. Cyber Comput. Commun.*, pp. 13–16, 2018.

[17] S. A. Jafar Zaidi, I. Chatterjee, and S. Brahim Belhaouari, "COVID-19 Tweets Classification during Lockdown Period Using Machine Learning Classifiers," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–8, Jul. 2022.

[18] V. Kandasamy *et al.*, "Sentimental analysis of covid-19 related messages in social networks by involving an n-gram stacked autoencoder integrated in an ensemble learning scheme," *Sensors*, vol. 21, no. 22, 2021.

[19] P. Sharma and T.-S. Moh, "Prediction of Indian election using sentiment analysis on Hindi Twitter," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 1966–1971.

[20] K. M. A. Hasan, M. S. Sabuj, and Z. Afrin, "Opinion mining using Naïve Bayes," in *2015 IEEE International WIE Conference on Electrical and Computer Engineering, WIECON-ECE 2015*, 2016, pp. 511–514.

[21] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake News Detection Using Sentiment Analysis," *2019 12th Int. Conf. Contemp. Comput. IC3 2019*, pp. 1–5, 2019.

[22] K. Poddar, G. B. D. Amali, and K. S. Umadevi, "Comparison of Various Machine Learning Models for Accurate Detection of Fake News," *2019 Innov. Power Adv. Comput. Technol. i-PACT 2019*, pp. 1–5, 2019.