Sentiment Analysis of COVID-19 Vaccines from Indonesian Tweets and News Headlines

Retnani Latifah
Teknik Informatika
Universitas Muhammadiyah Jaka
Jakarta, Indonesia
retnani.latifah@umj.ac.id

Ridwan Baddalwan
Teknik Informatika
Universitas Muhammadiyah Jaka
Jakarta, Indonesia
2017470105@ftumj.ac.id

Ambar Dwi Saputra
Teknik Informatika
Universitas Muhammadiyah Jaka
Jakarta, Indonesia
2017470080@ftumj.ac.id

Popy Meilina
Teknik Informatika
Universitas Muhammadiyah Jaka
Jakarta, Indonesia
popy.meilina@umj.ac.id

Yana Adharani
Teknik Informatika
Universitas Muhammadiyah Jaka
Jakarta, Indonesia
yana.adharani@umj.ac.id

***Abstract*—COVID-19 vaccines is a hot topic in online platforms due to the ongoing pandemic. Most studies on sentiment analysis of COVID-19 vaccines on Indonesian social media posts only used one or two classifiers with few modifications. This research investigated sentiment analysis using seven machine learning techniques on Twitter dataset which then used to predict on other unlabeled Twitter datasets as well as news headlines dataset. The experiment revealed that using the unigram bag-of-word KNN method gave the best result with 80.6% accuracy and slight overfit, 2.9% difference between training and testing accuracy. This classifier managed to capture a similar pattern of sentiment in Twitter datasets, which is dominated by neutral sentiment. In addition, the classifier could be used to classify different dataset, such as news headlines, with accuracy of 70% and could capture important words in negative sentiment.**

Keywords—sentiment analysis, COVID-19, vaccines, Twitter, news headlines

I. INTRODUCTION

The number of internet users and social media active users in Indonesia has increased (+15.5% and +6.3%), as of January 2021 [1]. One of the reasons is because of the on-going COVID-19 pandemic where people are advised to stay at home. Thus, most of their news is coming from either social media, online news outlets or other online platforms. During this time, people utilized social media as a platform for social interaction, collaboration, information sharing, learning, marketing and entertainment purposes [2]–[5].

These acts resulted in the growth of online posts, especially about COVID-19. One topic that has garnered a lot of attention since the pandemic started is the vaccines for COVID-19 virus. Since the beginning of the pandemic, news outlets have reported about any latest updates and social media users have also expressed their anticipations and concerns about this issue. From March 2020 to January 2021, the responses of Twitter users are gradually rising toward positive sentiment and their trust reached its peak in November 2020 when Pfizer announced their 90% effectiveness [6]. However, some people, including healthcare workers, are still hesitant about the vaccines mainly due to its safety [7]. Other reasons for the refusal, which were inferred from English-based tweets, are due to their fear and conspiracy theories [8].

Meanwhile, in Indonesian-language social media posts, the responses are varied, although the response for vaccine procurement is positive [9]. A research using tweets from August 25th to November 25th 2020 revealed that the sentiments were generally fluctuating, but when talking about vaccines produced by China, the sentiments are generally negative [10]. One similar result was shown from an analysis of the Ministry of Public Health Facebook Page [11].

However, a more neutral sentiment was found in research using 4941 tweets taken from October 25th to November 3rd, showing that tweets with general information are dominant [12]. Another research comparing the sentiments of Sinovac vaccine and Merah Putih vaccine (vaccines project supported by Indonesian Minister of Health) using tweets from November 2020 to February 2021 was done in [13] using Naive bayes and SVM. Both methods revealed that the sentiments are more inclined to be positive, which showed a shift of public opinion on China-produced vaccines from the previous study. Even so, in January when the president got his first injection the opinion was reserved because people did not trust the vaccines and feared its effect [14][15]. The polarity of Indonesian about COVID-19 keeps fluctuating as the vaccination program is in progress.

The previous studies of COVID-19 vaccines using Indonesian language-based posts are mostly done using one or two techniques such as Naive Bayes, SVM, lexicon based and Decision Tree. Few studies also did not clearly explain the procedure of the analysis. This paper analyzed the sentiment of COVID-19 vaccines from Indonesian-based tweets taken from March 23rd to 29th 2021 using seven machine learning techniques. The best model was picked to predict the rest of the unknown sentiments of Twitter dataset as well as news headlines. The addition of news headlines is because this research wishes to learn whether different text sources are able to classify the sentiment. The rest of the paper is organized as follows: Section II exhibits the methodology to analyze the data to get the sentiments. Section III discussed the results and findings and the relation of previous studies. The last one Section IV stated the conclusion and further research.

Funded by Pakarti Faculty of Engineering Universitas Muhammadiyah Jakarta

II. RESEARCH METHODOLOGY

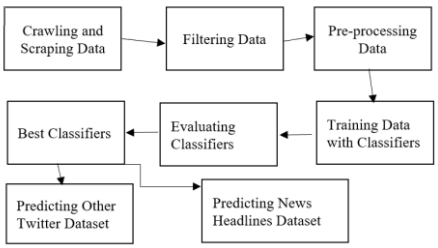


Fig. 1. Research Methodology

This research did a sentiment analysis about COVID-19 vaccines from Twitter dataset and news headlines. Sentiment analysis is a computational research that classifies text data according to its sentiment tendencies [16]–[18]. The methodology for this research, as seen in Fig. 1, consists of crawling and scraping both tweets and news headlines, filtering unnecessary and duplicated data, doing pre-processing data and modelled the training dataset. The best classifier was used to predict the rest sentiment of the dataset.

This study obtained 35000 tweets from March 23rd to March 29th 2021 using Tweepy. The keyword is "vaksin" (vaccine) and the language was set as 'id'. After deleting duplicate tweets and removing Malay language tweets, the final dataset is 9336 tweets. Among these, 6336 tweets were manually annotated. The other 3000 tweets were automatically labelled using the best classifiers. In addition, there are 6405 tweets which were randomly crawled in April and May and contain Malay language tweets with unknown sentiments. The example of annotated tweets is in Table I. The first tweet is an encouragement to get vaccinated, thus labelled as positive. The second tweet talked about the side effects of vaccines which scared the user to vaccinate their parents. The last tweet is a question, thus neutral.

TABLE I. EXAMPLE OF ANNOTATED TWEETS

tweet	label
jangan takut dan ragu untuk melakukan vaksinasi Covid-19 yaaa. Karena vaksin Covid-19 itu aman dan halal. #AyoVaksinTetap5M. Bangkit Indonesia Sehat	positive
Beberapa orang yang di vaksin di banjar, punya keluhan demam bahkan sampe ada yang muntah2. Hmm, jadi ragu ngizinin orang tua, tapi terlalu egois untuk ngelarang	negative
Jadwal vaksin buat rantau gimana ya	neutral

As for the news headlines dataset, it was scraped from antaranews.com, detikHealth and cnbcindonesia.com. Around 619 headlines were obtained. This data was automatically labelled using sentiment score, if the sum is more than 0 then annotated as positive, if less then negative and the rest are neutral [18].

All dataset underwent a preprocessing: tokenization, case folding, stopword removal and stemming using sastrawi python libraries[1)]. The stopword list was from Tala [19] with few additions such as 'yg', 'nah', 'nak', 'tu', and 'gak'. The example of manually annotated tweets dataset that has been preprocessed shown in Table II.

The data then split as training and testing set, and vectorized using unigram bag-of-words and TF-IDF methods. This dataset was modelled using several machine learning techniques as classifiers which are Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbor (KNN), Multinomial Naive Bayes (MNB), Random Forest Classifier (RFC), Decision Tree (DT) and Multilayer Perceptron (MLP). The best classifier was chosen by considering the accuracy on the test set and the possibility of overfitting. This classifier then used to automatically annotate the rest of the tweets and news headlines dataset. All the data that has been labelled with sentiments then being analyzed by investigating the word clouds from each sentiment.

TABLE II. EXAMPLE OF PRE-PROCESSING RESULTS

Original tweet	After pre-processing
b'Yuk, Jangan Takut dan Ragu Untuk Vaksinasi Covid-19\nKarena Vaksin Covid-19 Itu Aman dan Halal\n\nGuys, kamu jangan takut dan ragu untuk melakukan vaksinasi Covid-19 yaaa. Karena vaksin Covid-19 itu aman dan halal.\n #AyoVaksinTetap5M https://t.co/BoEJIDtVwz'	yuk takut ragu vaksinasi covid nkarena vaksin covid aman halal nguys takut ragu vaksinasi covid yaaa vaksin covid aman halal ayovaksintetap
b'Beberapa orang yang di vaksin di banjar, punya keluhan demam bahkan sampe ada yang muntah2. Hmm, jadi ragu ngizinin orang tua, tapi terlalu egois untuk ngelarang	orang vaksin banjar keluh demam sampe muntah hmm ragu ngizinin orang tua egois ngelarang
b'Jadwal vaksin buat rantau gimana ya \xf0\x9f\xa4\xa3'	jadwal vaksin rantau gimana ya

III. RESULT AND DISCUSSION

The result of sentiment analysis classifiers and the analysis of labelled datasets are as follow.

A. *Evaluation of Sentiment Analysis Classifiers*

The manually annotated tweets dataset has 3779 neutral tweets, 1839 positive tweets and 718 negative tweets. From the word cloud shown in Fig. 2 the neutral tweets were varied, astrazeneca, government policy or response and vaccination program. Some of these tweets came from news outlet accounts. Meanwhile, positive sentiments are represented by 'aman', 'halal' and 'vaksinasi' which represent invitations to get vaccinated. On the other hand, negative sentiments are about the conspiracy that vaccines contain pig substance which make it haram in Islam. Other notable words are about side effects and their fear of getting a vaccine.

Fig. 2. Wordcloud of Manually Annotated Tweets

The data was split into 80% training set and 20% testing set randomly. The training set was modelled using classifiers mentioned in the previous section. From the accuracy result on the test set shown in Fig. 3, all classifiers perform well on the data which is about 80%, except for Random Forest Classifier. However, many of them experienced overfitting except Random Forest Classifier, in both unigram bag-of-words and TF-IDF. Another classifier that performed well (80.6%) is KNN in unigram bag-of-words with only 2.4% difference of training and testing accuracy, followed by Multinomial Naive Bayes (2.9% difference). Logistic regression performed well in TF-IDF method (84.2% and 4.3% difference). The rest of the classifiers are consistent in both vectorizer methods, although the best accuracy is TF-IDF-SVM (84.7%). Unigram bag-of word-KNN was chosen as the best classifier and used to predict the rest of the dataset.

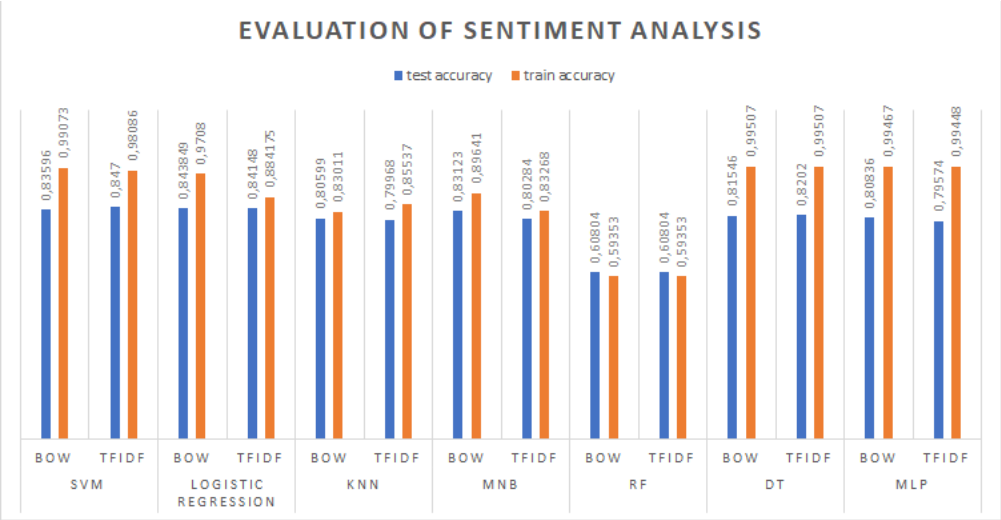


Fig. 3. Evaluation of Sentiment Analysis Classifiers on Twitter Dataset

B. *Predicting the Sentiment of Unlabelled Tweet Dataset*

The first dataset is 3000 tweets from March 2021. After removing duplicate tweets, 2554 tweets were predicted resulting in 219 positives, 46 negatives, and 2289 neutrals. The second dataset is tweets from April and May (6405 tweets after duplicate removal). Similar to previous data, neutral sentiment is dominating (5924), followed by 297 positive sentiments and 184 negative sentiments. This result is in-line with the report released by the Government of Jakarta. They did a sentiment analysis of vaccination in Jakarta from Twitter in May 2021. The result also indicates that most of the responses are neutral [20].

Tweets classified as neutral are dominating. One reason is because the manually annotated set is also dominated by neutral sentiments. The other sentiments also followed the distribution on the manually annotated set, positive sentiments are higher than negative sentiments. The same pattern also shown from word clouds, Fig. 4. From the word cloud of March dataset, the positive sentiment is about encouragement to get vaccinations to fight the pandemic, which is part of the government campaign. Meanwhile, the negative sentiment is slightly different where conspiracy about vaccines containing pig substance is not present and only words about side effects and fear.

The second dataset was taken around Ramadhan month and Eid Mubarak celebration, which means people are moving from one province to another and even to different islands. Thus, in the positive sentiment, from the word cloud in Fig. 4, there is also a suggestion to not 'mudik (going back home)' to prevent massive transmission. In negative sentiments, the discussion about pig substance is back, although not as massive as in March.

C. *Predicting the Sentiment of News Headlines Dataset*

The news headlines dataset was automatically annotated using sentiment scores with 466 neutrals', 126 positives' and 27 negatives'. The unigram bag-of-word KNN classifier also gives similar results (564, 45 and 10). The accuracy of the two labels is 70%, but the macro average F1-measure is only 31% and the weighted average F1-measure is 64%. This means that the model actually did not perform quite well on minority classes toward the sentiment score label.

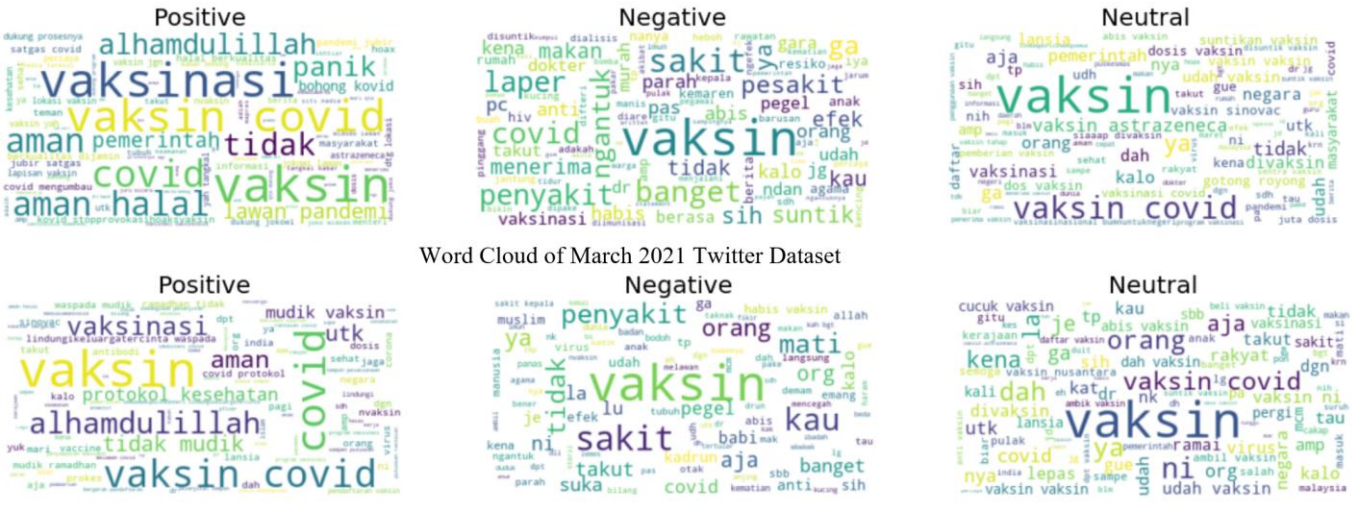
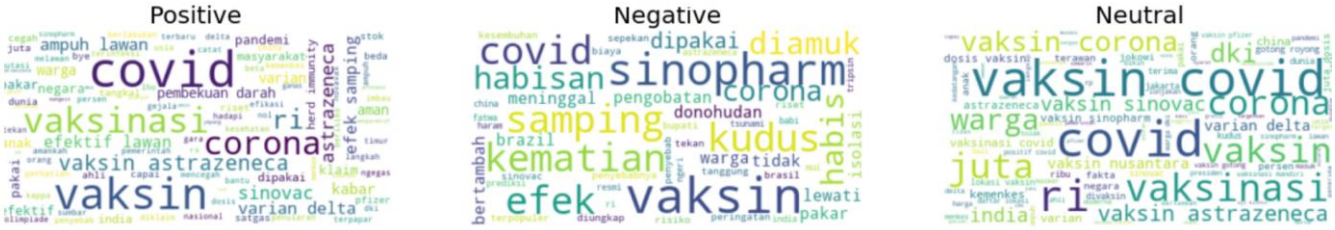


Fig. 4. Evaluation of Sentiment Analysis Classifiers on Twitter Dataset

From the captured word cloud, seen in Fig. 5, both are different yet similar. In negative word clouds for example, both are able to capture 'kematian (dead)' but sentiment score labels could not capture the word 'kudus' and 'bertambah (increase)'. These two words are important because at the time there are a surge of COVID-19 cases in Kudus, which made as headlines. This means that KNN classifiers might be slightly better at analyzing the sentiment of news headlines. Overall, it can be said that the classifier which is based on the Twitter data model is able to classify another type of dataset. Although, there is a need to further investigate if the news headlines get a proper annotation.

Word Cloud of News Headline Sentiment Score Label



Word Cloud of News Headline Unigram KNN Label

IV. Conclusion and Further Research

This research investigated the sentiment of the public about COVID-19 vaccines. While, some previous study already did the work, this study took different Twitter dataset, compared several classifiers technique and also used the model to classify not only the unlabeled Twitter datasets but also news headlines datasets about COVID-19 vaccines. The best classifier chosen is unigram bag-of-word KNN method with 80.6% accuracy on the test set and only has slight overfit (2.9%).

The classifier is able to classify both datasets with similar pattern to the validation set for each sentiment, mostly neutral followed by positive and a few negative. The classifier is also run-on news headlines dataset which has 70% accuracy but only 31% macro average F1-measure and 64% weighted average F1-measure. From the word cloud, the classifier might be slightly better in capturing important words.

The plan for future research is to use Indonesian BERT model to pre-trained the dataset. In addition, to better capture the topic of each sentiment, the use of topic modelling methods such as Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) might be good to cluster the topics.

References

[1] S. Kemp, "Digital in Indonesia: All the Statistics You Need in 2021 — DataReportal – Global Digital Insights," *Dataeportal*, 2021. https://datareportal.com/reports/digital-2021-turkey%0Ahttps://datareportal.com/reports/digital-2021-indonesia?rq=indonesia.

[2] A. Wong, S. Ho, O. Olusanya, M. V. Antonini, and D. Lyness, "The use of social media and online communications in times of pandemic COVID-19," *J. Intensive Care Soc.*, vol. 22, no. 3, pp. 255–260, 2020, doi: 10.1177/1751143720966280.

[3] H. Junawan and N. Laugu, "Eksistensi Media Sosial,Youtube, Instagram dan Whatsapp Ditengah Pandemi Covid-19 Dikalangan Masyarakat Virtual Indonesia," *Baitul 'Ulum J. Ilmu Perpust. dan Inf.*, vol. 4, no. 1, pp. 41–57, 2020, doi: 10.30631/baitululum.v4i1.46.

[4] M. A. Harahap and S. Adeni, "Tren penggunaan media sosial selama pandemi di indonesia," *J. Prof. FIS UNIVED*, vol. 7, no. 2, pp. 13–23, 2020.

[5] N. N. Rohmah, "Media Sosial Sebagai Media Alternatif Manfaat dan Pemuas Kebutuhan Informasi Masa Pandemik Global Covid 19 (Kajian Analisis Teori Uses And Gratification)," *Al-I'lam J. Komun. dan Penyiaran Islam*, vol. 4, no. 1, pp. 1–16, 2020, [Online]. Available: https://journal.ummat.ac.id/index.php/jail/article/view/2957/1905.

[6] J. C. Lyu, E. Le Han, and G. K. Luli, "COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis," *J. Med. Internet Res.*, vol. 23, no. 6, p. e24435, 2021, doi: 10.2196/24435.

[7] A. Richardson, "Health Care Workers' Reluctance to Take the Covid-19 Vaccine: A Consumer- Marketing Approach to Identifying and Overcoming Hesitancy," *Biomedgerontology*, no. 631, pp. 1–10, 2020, doi: 10.1056/CAT.20.0676.

[8] M. Thelwall, K. Kousha, and S. Thelwall, "Covid-19 Vaccine Hesitancy on English-Language Twitter," *Prof. la Inf.*, vol. 30, no. 2, pp. 1–13, 2021, doi: 10.3145/epi.2021.mar.12.

[9] M. I. Aditama, R. I. Pratama, K. H. U. Wiwaha, and N. A. Rakhmawati, "Analisis Klasifikasi Sentimen Pengguna Media Sosial Twitter Terhadap Pengadaan Vaksin COVID-19," *J. Inf. Eng. Educ. Technol.*, vol. 4, no. 2, pp. 90–92, 2020.

[10] M. Hidayat, "Kupas Data_ Vaksin Covid-19, Antara Harapan dan Keraguan - Tekno Liputan6," *Liputan6.com*, 2020.

[11] A. Harun and D. P. Ananda, "Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve Bayes dan Decission Tree," *Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. April, pp. 58–63, 2021.

[12] F. F. Rachman and S. Pramana, "Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada

Media Sosial Twitter," *Heal. Inf. Manag. J.*, vol. 8, no. 2, pp. 100–109, 2020, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/223/175.

[13] C.- Pandemic, B. Laurensz, and E. Sediyono, "Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 2, pp. 118–123, 2021.

[14] 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," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1088, no. 1, p. 012045, doi: 10.1088/1757-899x/1088/1/012045.

[15] G. Y. Pratama, C. A. Udin, D. A. Aprilian, M. R. Erliansyah, and F. Rumaisa, "Public Sentiment Towards The COVID 19 Vaccine in Indonesia," *Turkish J. Physiother. Rehabil.*, vol. 32, no. 3, pp. 6081–6084, 2021.

[16] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *J. Big Data*, vol. 5, no. 1, pp. 1–10, 2018, doi: 10.1186/s40537-018-0164-1.

[17] T. N. Fatyanosa and F. A. Bachtiar, "Classification method comparison on Indonesian social media sentiment analysis," *Proc. - 2017 Int. Conf. Sustain. Inf. Eng. Technol. SIET 2017*, vol. 2018-Janua, no. April, pp. 310–315, 2018, doi: 10.1109/SIET.2017.8304154.

[18] E. B. Santoso and A. Nugroho, "Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik Di Facebook," *Eksplora Inform.*, vol. 9, no. 1, pp. 60–69, 2019, doi: 10.30864/eksplora.v9i1.254.

[19] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.

[20] JSC, "Analisis Sentimen Vaksinasi Covid-19 di Jakarta," 2021. [Online]. Available: https://smartcity.jakarta.go.id/blog/729/analisis-sentimen-vaksinasi-covid-19-di-jakarta.