icimcis

by Retnani Latifah

Submission date: 08-Mar-2022 11:40AM (UTC+0700)
Submission ID: 1779159807
File name: and_News_Headlines_using_Various_Machine_Learning_Techniques.pdf (402.95K)
Word count: 3991
Character count: 22095

Sentiment Analysis of COVID-19 Vaccines from Indonesian Tweets and News Headlines using Various Machine Learning Techniques

1st Retnani Latifah
Informatics Engineering
Universitas Muhammadiyah Jakarta
Jakarta, Indonesia
retnani.latifah@umj.ac.id

2nd Ridwan Baddalwan
Informatics Engineering
Universitas Muhammadiyah Jakarta
Jakarta, Indonesia
2017470105@ftumj.ac.id

3rd Popy Meilina
Informatics Engineering
Universitas Muhammadiyah Jakarta
Jakarta, Indonesia
popy.meilina@umj.ac.id

4th Ambar Dwi Saputra
Informatics Engineering
Universitas Muhammadiyah Jakarta
Jakarta, Indonesia
2017470080@ftumj.ac.id

5th Yana Adharani
Informatics Engineering
Universitas Muhammadiyah Jakarta
Jakarta, Indonesia
yana.adharani@umj.ac.id

2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS) | 978-1-6654-2733-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIMCIS53775.2021.9699187

***Abstract*—COVID-19 vaccine 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 in which the one with the highest evaluation value will be used to predict on other unlabeled Twitter datasets as well as news headlines dataset. The same classifier is also used to build a visualization dashboard that reflect the result of the sentiments. The result from the sentiment classification is then used to identify the topics, by using word cloud. The experiment revealed that SVM classifier has the highest accuracy and micro average F1-measure, which is 84% and 0.76. This classifier managed to capture similar patterns of sentiments in Twitter and news headlines datasets, which is dominated by neutral sentiment. Some of the topics from each sentiment, managed to reflect the real condition when the datasets were collected.**

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. 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 and compared the result. The model with the highest evaluation performance. 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.

II. Research Methodology

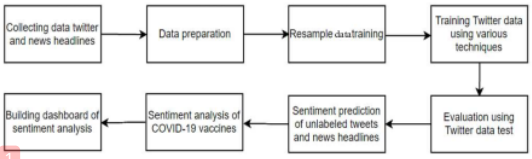

Fig. 1. Research Methodology

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, is started by collecting the data from twitter and Indonesian news headlines from March 23rd to March 29th 2021 using Tweepy. The keyword used is "vaksin" (vaccine) and the language was set as 'id'. After deleting duplicate tweets and removing some Malay language tweets, the final Twitter dataset has 6336 tweets. Another 9405 tweets are collected on April and May. 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].

The first dataset is manually annotated and used to make model for predicting the sentiments. The example of annotated tweets is in Table I. The first tweet is an encouragement to get vaccinated, thus labeled 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.

All dataset underwent a preprocessing such as lowercasing, noise removal such as digit and symbols, 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 annotated dataset has three label which are positive (1839), negative (718), and neutral (3779). Which then split into 80% training consist of 3008 neutrals, 1471 positives and 589 negatives. The rest of the annotated dataset is used as testing dataset. The SMOTE oversampling technique is used to balanced the training dataset. Because there are two minority labels, the oversampling is done two times, thus each label has 3008 tweets. Both the training and testing dataset then vectorize using bag-of-words and TF-IDF methods in NLTK library.

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

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 Co 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

The training dataset then model using seven machine learning techniques namely 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). This study used Sklearn library to generate the model from each classifier. The parameters used are the default setting.

The algorithms then evaluated using confusion matrix to get one classifier with the highest macro average F1-measure and accuracy. This classifier will be used to predict the unlabeled Twitter dataset and news headlines and build a visualization dashboard of the sentiment result. The tweets and news headlines that have been labeled then analyze using word cloud to get the topic of each sentiment.

III. Result and Discussion

A. Evaluation of Sentiment Analysis Classifiers

This study experimented with two vectorizer, bag-of-words and TFIDF. The result from using bag-of-words vectorizer showed that the value of macro-average F1 measure is below or equal 0.7, as seen in fig 2. Classifier with 0.7 is only multinomial naive bayes. However, the figure also showed that there are four classifiers with more than 70% accuracy. Compared to previous study, this result is not as good which get 85% [20] and 100% [11] in accuracy. This is might be because there is a difference in dataset as well as the labelling process. Because this study, the labelling is done manually, there might be a different step compared to previous studies.

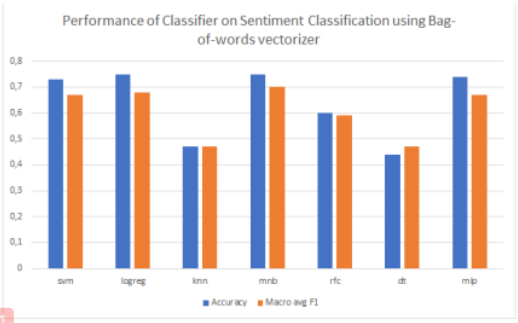

Fig. 2. Performance of Classifier on Sentiment Classification using Bag-of Words Vectorizer

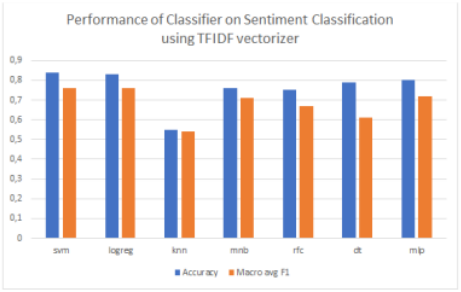

Fig. 3. Performance of Classifier on Sentiment Classification using TFIDF Vectorizer

Meanwhile using TFIDF vectorizer, the accuracy of SVM and Logistic Regression are 84% and 83% with macro-average F1 measure for both classifiers are 0.76. Compared to other classifiers these two algorithms have the highest evaluation result. Both are linear model using several iteration processes to find the optimal model. Meanwhile, kNN algorithm in both vectorizer showed a low performance compared to the others. This maybe because kNN could not make the right model using the distance of each term.

From the evaluation result, SVM is chosen as the classifier to predict the unlabeled tweets and news headlines because the accuracy is higher compared others. SVM is able to label the unlabeled Twitter dataset into three sentiments, as seen in Fig 4 (left figure), in which most tweets are labelled as neutral (6999). Neutral label is colored as orange, while negative is green and positive is blue. From the result, it could be seen that SVM assigned more negative label compared to positive although the original training dataset has more positive label, which might due to oversampling process.

However, oversampling did not affect the classification result that most of the tweets about vaccines in Indonesian language are neutral. This 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 [21]. This showed that during the time of data collection, most Indonesian did not have particular complaints or praises about vaccination program. When the data was collected, the pandemic situation was stabilizing although the cases were still high.

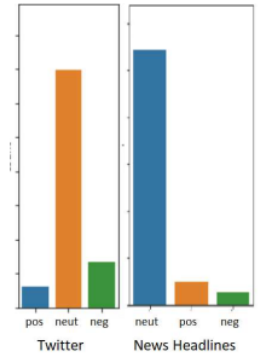

Fig. 4. Label Distribution from SVM Classifier on Twitter Unlabeled Dataset and News Headlines

Fig. 5. Wordcloud of Manually Annotated Tweets

For news headlines dataset, the result is similar to Twitter dataset which is more data are assigned as neutral (542). However, the second one is positive and the least assigned label is negative. Journalist should pick neutral words for news headlines, thus this result showed that most of news headlines about COVID-19 vaccines in the dataset followed this attitude. However, there are 27 data that are labelled as negative which might be because there are words that are mostly associated with negative label in the training dataset that are used in headlines.

B. Analysis of Sentiments

From the word cloud of manually annotated dataset shown in Fig. 5 the topic in the neutral tweets were varied such as astrazeneca, sinovac, distribution of vaccine and its dosage. This means that most of the tweets most likely are about information and not expressing opinion. Tweets from news outlet or organization are also dominant. Meanwhile, positive sentiments are represented by 'aman', 'halal' and 'vaksinasi' which represent invitations to get vaccinated. On the other hand, tweets with negative sentiments are about the conspiracy that vaccines contain pig substance which make it haram in Islam. The words are 'babi' and 'haram'. Other notable word is about the fear of getting a vaccine.

Meanwhile from the second Twitter dataset labelled using Support Vector Machine is not filtered and there are tweets which has Malay language and assigned as neutral label. However, on positive label the topics are similar to the manually annotated dataset. The identified topic are about the halal status of COVID-19 vaccines, campaign to not travel for Ied Mubarak holiday by the government (word 'tidak mudik'), and campaign to get vaccination and to do the preventive measurements ('protokol kesehatan') so people can protect their own family ('lindungikeluarga). This result reflected the real situation when the data was collected. As for the negative label, the topics are similar to the manually annotated dataset used in modelling phase. Meanwhile, the neutral label is not focused on certain topics, showed by no particular words are highlighted beside 'vaksin' and 'covid'. In this label, it also showed that the preprocessing phase need to be improve because there are words that should be remove but still present. This is also because there are Malay tweets which has not been filtered from the unlabeled Twitter dataset.

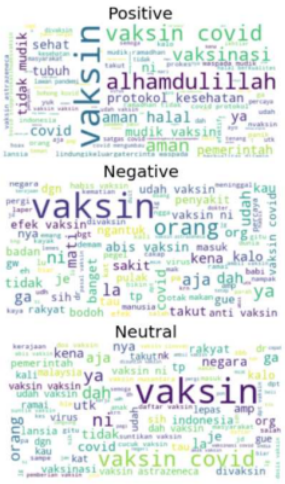

Fig. 6. Wordcloud of Automatically Labelled Twitter Dataset using SVM

The topics of news headlines with positive label from the word cloud are about the current situation of the type of vaccines. While reporting about side effect of vaccine which is one of it is blood clot (showed by 'efek samping' and 'pembekuan darah'), some headlines use positive related words used in training dataset. Another topic identified from positive label is about the effectiveness of vaccines, showed by 'efektif', 'ampuh' and 'lawan'. Meanwhile in negative label, the topics are about vaccines from China such as sinovac and sinopharm, the situation of pandemic in india, about the former minister 'terawan' who was claimed by many that he failed to control the pandemic since the beginning, about nusantara vaccine which has been developed locally and about death due to COVID. For neutral label, the notable words beside 'covid' and 'vaksin' is 'varian delta' which reported about the new emerging and deadly variant.

From these topic identifications based on its sentiment on news headlines dataset, result showed that SVM classifier is able to correctly assigned the sentiment label for positive and negative, although the training dataset is limited.

Fig. 7. Wordcloud of Automatically Labelled News Headlines Dataset using SVM

The topic is also detected using Non-negative Matrix Factorization. From the Twitter dataset, the top topic is about information of vaccination program with 'suntik', 'juta' and 'dosis'. The second one is about opinion on vaccine such as using the word 'aman', 'takut', 'ragu'. The third topic is about the encouragement of not travelling, get vaccination and protect family, noted by 'tidak', 'mudik', 'lindungi', 'ayo', and 'dukung'

In news headlines dataset, the first topic is about type of vaccines such as 'astrazeneca', 'sinovac', 'nusantara' and 'sinopharm' in correlation to vaccination program. The second topic is about positive cases in India which is increasing drastically due to delta varian. The word from this topic included 'india', 'positif', 'tambah' and 'lonjak'. The third topic is about delta variant which has entered Indonesia. Beside 'delta' and 'varian', other notable word is 'Kudus' which is at the time of the data collection the city had a surge of positive patient due to delta variant.

Overall, the topics of both Twitter and News headlines dataset are similar and reflected the real situation at the time the data were taken. The automatic labelling is also doing fairly well in classifying the sentiments and could be used not only in unlabeled Twitter data but also in news headlines data.

C. *Build Dashboard*

The SVM model is used to build a dashboard app in which a user could see the visualization of sentiments distribution and frequent username from the Twitter dataset. There is a need to improve the dashboard so it could be used for in predicting the sentiments of new text, so it could be more practical. The image of the dashboard can be seen in Fig 8.

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 the one with highest accuracy and macro average F1-measure. The classifier chosen is Support Vector Machine with 84% accuracy and 0.76 macro-average F1-measure.

The model using SVM classifier is able to classify both unlabeled Twitter and news headlines dataset. Neutral sentiment is dominating in both dataset due to the training dataset also dominated by neutral.

Fig. 8. Dashboard of SVM Model

However, in Twitter dataset, more data is assigned as negative compared to positive although in the original dataset,

the positive data is higher. This is may be due to the oversampling technique implemented in training dataset. Meanwhile, in news headlines dataset, the distribution of the label is similar to training dataset which are mostly neutral as news headlines should be doing, followed by positive and negative label.

The topics detected from each label showed similar pattern, either in Twitter dataset and news headlines dataset. Positive label is about the safety of vaccines, and also government campaign about vaccination, stay at home and protect their own family. Meanwhile in negative label is about the increase of positive cases due to delta variant.

In the future there is a need to refined the pre-processing data and labelling the training dataset. The dashboard app also needs to get improvement so it could be used by public conveniently.

ACKNOWLEDGMENT

This publication is a part of a research grant funded by Faculty of Engineering Universitas Muhammadiyah Jakarta Indonesia, through PAKARTI.

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] 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] A. L. Fairuz, R. D. Ramadhani, and N. A. Tanjung, "Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial," *J. DINDA*, vol. 1, no. 1, pp. 10–12, 2021, [Online]. Available: http://journal.ittelkom-pwt.ac.id/index.php/dinda/article/view/180.

[21] 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.

icimcis

ORIGINALITY REPORT

94%	3%	94%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Retnani Latifah, Ridwan Baddalwan, Popy Meilina, Ambar Dwi Saputra, Yana Adharani. "Sentiment Analysis of COVID-19 Vaccines from Indonesian Tweets and News Headlines using Various Machine Learning Techniques", 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021 Publication	94%

Exclude quotes	On	Exclude matches	< 3%
Exclude bibliography	On		