SENTIMENT ANALYSIS OF INDONESIAN COMMUNITY ON COVID-19 VACCINATION ON TWITTER SOCIAL MEDIA

Widi Astuti[1]; Windu Gata[2]; Nurmalasari[3]; Ida Zuniarti[4]

[1] Digital Business, [2] Informatics, [3] Information Systems, [4] Management Study Program
Unviersitas Nusa Mandiri
www.nusamandiri.ac.id
[1] widiastuti.wtu@nusamandiri.ac.id, [2] windu@nusamandiri.ac.id,
[3*)] nurmalasari.nmr@nusamandiri.ac.id, [4] ida.idz@nusamandiri.ac.id

(*) Corresponding Author

Abstract—In the process, data mining will extract valuable information by analyzing the existence of specific patterns or relationships from extensive data. One of the concerns of the new disease outbreak caused by the coronavirus (2019-nCoV) or commonly referred to as Covid-19, was officially designated as a global pandemic by the World Health Organization (WFO) on March 11, 2020. To break the transmission of Covid-19, the government carried out vaccinations for the Indonesian population. In the first period, the vaccination target will be for health workers with a total of 1.3 million people, public officers with 17.4 million people, and 21.5 million people. 19. The Data processed is only text data from Twitter application reviews that use Indonesian. Using the polarity of the Sentiment class Textblob, the sentiment class is positive, negative, and neutral. The data mining used is SVM, Naive Bayes, and Logistic Regression. As for this research in the form of knowledge of sentiment in the community towards vaccination activities, the results of this study get 43% positive sentiment, 40.8% negative, and 16.2% negative by testing the classification algorithm, Logistic Regression accuracy of 87%, SVM 86, 4%, and Naive Bayes, 40% of these results, can be seen that the Indonesian people have a positive sentiment towards the covid-19 vaccine.

Keywords: data mining, covid-19 vaccine, Twitter, naive Bayes, SVM, logistic regression.

Abstrak—*Dalam prosesnya data mining akan mengekstrak informasi yang berharga dengan cara menganalisis adanya pola-pola ataupun hubungan keterkaitan tertentu dari data- data yang berukuran besar. Salah satu perhatian wabah penyakit baru yang disebabkan oleh virus korona (2019-nCoV) atau yang biasa disebut dengan Covid-19 yang di tetapkan secara resmi sebagai pandemi global oleh World Health Organization (WFO) pada tanggal 11 Maret 2020. Dalam rangka memutus penularan Covid 19 pemerintah melakukan tindakan vaksinasi kepada penduduk indonesia. Pada periode pertama target vaksinasi yang akan di lakukan yaitu kepada tenaga kesehatan dengan jumlah 1,3 juta orang, petugas publik 17,4 juta orang dan 21,5 juta orang, penelitian ini mengambil data dari ulasan pengguna aplikasi Twitter dengan topik pembahasan vaksin covid-19. Data yang diolah hanya data teks dari ulasan aplikasi Twitter yang mengguanakan bahasa Indonesia. Menggunakan polaritas dari Textblob kelas sentimen yang digunakan kelas sentimen positif, negatif dan netral. Data mining yang digunakan adalah SVM, Naive Bayes dan Logistic Regression. Adapun dari penelitian ini berupa pengetahuan sentimen di tengah masyarakat terhadap kegiatan vaksinasi, hasil dari penelitian ini mendapatkan 43% sentimen Positif, 40,8% Negatif dan 16,2% negatif dengan pengujian terhadap algoritma klasifikasi akurasi Logistic Regression sebesar 87%, SVM 86,4%, dan Naive Bayes 40% dari hasil tersebut dapat dilihat bahwa masyarakat indonesia bersentimen positif terhadap vaksin covid-19.*

Kata kunci*: data mining, vaksin covid-19, Twitter, naive Bayes, SVM, logistic regression.*

INTRODUCTION

Technology has become a significant need in education, one of which is data mining technology used in research. Data mining will extract valuable information by analyzing specific patterns or relationships from extensive data. Several surveys on the modeling process and methodology stated, "Data mining is used as a guide, where data mining provides a summary of the history, description and as a standard guide regarding the future of a data mining model process (Nur Khormarudin, 2016). One of the concerns of the new disease outbreak caused by the coronavirus (2019-nCoV) or commonly referred to as Covid-19, was officially designated as a global pandemic by the World Health Organization (WFO) on March 11, 2020, ago (Rachman & Pramana, 2020).

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 2 September 2022 | DOI: 10.33480/pilar.v18i1.3820

More than 41.5 million cases and more than 1.1 million deaths are expected to occur worldwide by October 23, 2020, with the Chinese city of Wuhan serving as the epicentre of the virus's spread at the end of 2019. To stop the transmission of Covid 19, the government has carried out vaccinations for the Indonesian population. In the first period, the vaccination targets to be carried out are 1.3 million health workers, 17.4 million public officers, and 21.5 million older adults (Yolanda, 2021). With this step, it is hoped that cases of infection with the virus in Indonesia will improve soon and people can carry out their daily activities as before.

Vaccines are the most efficient and cost-effective way of preventing infectious diseases. The Minister of Health has determined seven types of Covid-19 vaccines to be used for vaccination in Indonesia (Dewi, 2021). The types of vaccines that can be used are those produced by PT. Bio Farma, AstraZeneca, Sinopharm, Moderna, Novavax, Inc., Pfizer Inc, and BioNTech and Sinovac (Susilo Daniel & Navarro, 2021). Some vaccines have advantages and disadvantages and have different effects on the human body. Indonesia is an internet user based on sources from We are social. In 2020, Indonesia reached 175.4 million people with internet access. Twitter is one of the most widely used media platforms in Indonesia. According to sources from We Are Social and Hootsuite, in 2020, Twitter ranked fifth in the most often used category, with a comprehensive presentation of 56% after Youtube, Whatsapp, Facebook, and Instagram (Keahlian & Data, 2021). The use of Twitter features to write thoughts and opinions by platform users (K-means, 2017). These different opinions are significant and one of the things that influence the primary human behavior to get results from sentiment. Based on the current Twitter trend (Ahuja & Dubey, 2017), research on COVID-19 is fascinating to examine to find out public opinion about vaccination(Makmun & Hazhiyah, 2020) during the ongoing pandemic, both positive, negative, and neutral responses.

MATERIALS AND METHODS

This research uses data from Twitter in the form of tweets from Twitter application users with the topic of discussion regarding the covid-19 vaccine. The Data that has been crawled is 12,150 from April 19, 2021, to May 6, 2021. this research tool utilizes the google collaboration menu with the python programming language(ian h. written, eibe frank, mark a. hall, 2017). Moreover, in the model in the labeling process, the researcher uses the textblob method to categorize the sentiments of positive, negative, and neutral, while the classification methods used are logistic regression (Primadhita & Budiningsih, 2020), SVM, and naive Bayes.

The process steps used in this research are cross-industry standard process for data mining (crisp-dm) which is an approach framework for translating business problems into data mining tasks and carrying out data mining projects that are separate from the application area and technology used (mart, Contreras-Machado, & machine, 2019).

the stages of the crisis phase are as follows (Kurniawan et al., 2019):

a. Business Understanding

Every data mining project begins with defining project objectives, including the first phase, business understanding. The business aims to maximize uptime and machine efficiency by using predictive analytics. This target is then converted into data mining by identifying the relevant machine components.

b. Data Understanding

Data mining project objectives are formed based on experience and qualified assumptions. In the Data Understanding phase, information about predictive maintenance scenarios is hidden to detect faults, a valid concept to look for new frequency patterns in the data stream of a sensor movement.

c. Data Preparation

In the Data Preparation phase, the researcher collects relevant data and prepares data mining that uses preprocessing, such as data reduction, filtering, and feature creation related to the objectives of the data mining project.

d. Modelling

In the Modeling phase of data mining, workflows are built to find the desired parameter settings and selected algorithms to execute. The task of data mining is on the previously processed data.

e. Evaluation

In the Evaluation phase, testing the model against real data sets in production scenarios and assessing data mining results based on business goals. For this purpose, generate a test data set following the steps developed in the "Data Preparation" and "Modeling" phases, excluding the labelling step.

f. Deployment

After a successful evaluation, use the training model in production in the Deployment phase. Its deployment requires a stable setup for data acquisition, including data processing infrastructure.

RESULT AND DISCUSSIONS

The following are the stages of the research phase carried out:

A. Business Understanding

At this stage, an understanding of the object of research by finding information through social media Twitter about the ongoing Covid-19 vaccination in Indonesia, expressing various kinds of opinions, both negative and positive, on tweets of social media users.

Implementation of business understanding helps determine the best sentiment analysis approach and a suitable model based on a comparison of algorithm results. The algorithms used are Logistic Regression and SVM. The following figure 1 is the framework of this study:

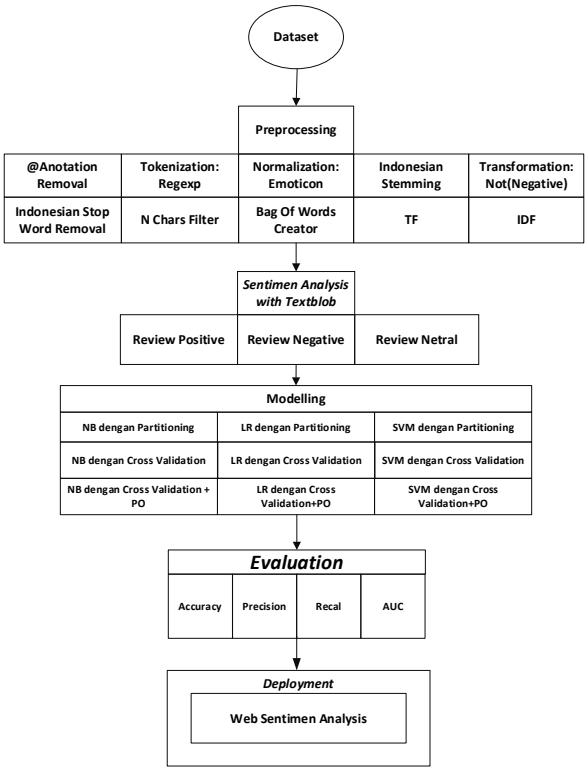

Figure 1. Thinking Framework

A. Data Understanding

The Data Understanding stage aims to collect, identify, and understand the data held. The data must also be verifiable. The data used in this study is data from Twitter user reviews in response to the news of the Covid-19 Vaccine, which consists of positive and negative categories. Table 1 contains Twitter data about the Covid-19 vaccine.

Table 1. Tweets about the Covid-19 Vaccine

Date	*Tweet*
April 19, 2021	*talking about vaccines* male w: what is a good vaccine? w: pfizer? sinovac? genose? male w: no, there are 2 words w: what is it... male w: there is... what is.... there is astra... mybos? THERE IS GENSHIN'"
April 20, 2021	To receive the second dose is 12 weeks after the first injection for the Astrazeneca vaccine type. Work Hard Forward My Country
April 22, 2021	Officially, Denmark Stops Using the Covid-19 Vaccine from AstraZeneca. The decision was taken after many findings of rare cases of blood clots in Danes after being injected with the vaccine. * Dangerous Viruses or Vaccines?!
April 29, 2021	The point is, even though you've been vaccinated, it does not mean you're free from covid.. Vaccination is just the first step, so do not be easily fooled by news headlines Vaccination is running, the positive number of covid is decreasing and the next day you gather together with friends in a small amount

B. Data Preparation

The Data Preparation stage was obtained from crawling data using the Tweepy library with the Python programming language from reviews on the Twitter application for the Covid-19 vaccination keyword, consisting of positive and negative categories. The data cleansing to reduce duplicate and redundant data, then tokenizing, word normalization, stopword removal, and stemming.

1. @#Annotation removal: Regexp

The @#Annotation removal process is the process of removing text that has @ and # annotations. This preprocessing process is done using python using the regular expression command re.sub ("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text). So that tweets with user mentions and hashtags disappear. In Table 2, Data before and Data after in Data Annotation Removal.

Table 2. Data Annotation Removal

@#Annotation Removal	
Data before	Data after
#Repost @puskesmas_kalitanjung @download.ins The first stage of vaccination for the elderly at RW 03 Harjamukti Kota Cirebon.	First stage of elderly vaccination RW 03 Harjamukti Kota Cirebon.

2. Tokenization: Regexp

Tokenization process: Regexp is a process to remove punctuation marks and numbers so that the result is a word. This preprocessing process out using the Regexp library in python, in table 3 Data before and Data after when tokenization process.

Table 3. Tokenization Process

Tokenization: Regexp	
Data before	Data after
Have you been vaccinated yet, friends?	Have you been vaccinated yet, friend

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 2 September 2022 | DOI: 10.33480/pilar.v18i1.3820

3. Indonesian Stemming

The Indonesian Stemming process is a process to find the root word of a word. This preprocessing process uses the Python Sastrawi library. Table 4 Data before and Data after when the process of Indonesian Stemming.

Table 4. Indonesian Stemming

Indonesian Stemming	
Data before	Data after
This material is also imported from the USA, it is the same. After all, Sinovac, whose imported ingredients are made in Bandung, is also the Nusantara vaccine, so what happens	This material is also imported from the USA, it's the same, after all, Sinovac, which is imported, makes Bandung and the nusantara vaccine, so please

4. Indonesian Stop Word Removal

The Indonesian Stop word removal process is a process to remove common words that usually appear in large numbers and are considered meaningless. Some examples are yg, dg, rt, dgn, ny, d, klo, kalo, amp, biar, bikin, bilang,gak, ga, krn, nya, nih, sih, si, tau, tdk, tuh, utk, ya, etc. Table 5 Data before and after the Indonesian Stop Word Removal process.

Table 5. Indonesian Stop Word Removal

Indonesian Stop word removal	
Data before	Data after
In the beginning, the vice president, Kh Maaruf Amin, refused the Sinovac vaccine and waited for the Pfizer vaccine.	The poem of the vice president, kh maaruf amin, rejects the sinovac vaccine, waits for the pfizer vaccine, the dynamics of the country's vaccine embargo, the spread of covid increases

5. N Chars Filter

The N Chars Filter process is to remove words that are less than one syllable. In table 6 is Data before and Data after N.Chars Filter.

Table 6. N.Chars Filter

N Chars Filter	
Data before	Data after
I want Sinovac, but if Astra is back, it's better to vaccinate another route, byr GPP as long as you don't have Astra Zeneca	I want Sinovac, but if my Astra is back, it's better to pay for the vaccine from another route, it's okay as long as it's not Astra Zeneca

6. Textblob labelling results

In this process, it calculates the sentiment of each tweet to process textual data. At this stage, using the Natural Language Processing (NLP) technique and tweeting data generally belongs to Unsupervised Learning. NLP is required to identify opinions and sentiments and classify them into positive, negative, or neutral labels. The python library used to identify sentiment analysis in this study is Textblob. Moreover, from each of these tweets, the polarity is determined whether it means positive, negative, or neutral. Table 7 is the result of Labeling Textblob consisting of sentiment, count, and percentage.

Table 7. Results of Labeling Textblob

Sentiment	Count	Percentage
Positive	3.431	43,0%
Neutral	3.259	40,8%
Negative	1.296	16,2%
Total	7.986	100%

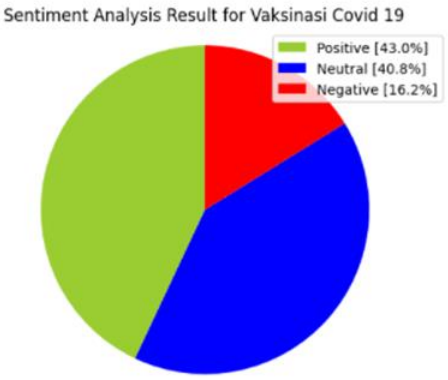

Figure 2. Results of Labeling Textblob

In figure 2, view percentages in charts for the result of labelling textblob.

C. Modelling

The classification technique for this modelling stage is by comparing three algorithms: naïve Bayes, logistic regression, and support vector machine. Table 8 is a script for the naïve Bayes model, table 9 is the logistic regression model, and table 10 is the script for the SVM model.

1. Modelling Naive Bayes

Table 8. Script Modelling Naive Bayes

```
from sklearn.naive_bayes import GaussianNB
modelnb = GaussianNB()
nbtrain = modelnb.fit(X_train, y_train)
y_pred = nbtrain.predict(X_test)
acc = accuracy_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred,average='micro')
cr = classification_report(y_test, y_pred)
```

2. Modelling Logistic Regression

Table 9. Script Modelling LR

```
from sklearn.linear_model import LogisticRegression
clf_lr = LogisticRegression(C = 1.2)
clf_lr.fit(X_train, y_train)
y_pred = clf_lr.predict(X_test)
acc = accuracy_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred,average='micro')
cr = classification_report(y_test, y_pred)
```

3. Modelling SVM

Table 10. Script Modelling SVM

```
from sklearn.svm import SVC
clf_svm = SVC(kernel="linear", C=1)
clf_svm.fit(X_train, y_train)
y_pred = clf_svm.predict(X_test)
acc = accuracy_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred,average='micro')
cr = classification_report(y_test, y_pred)
```

D. Evaluation

The comparison of the results of accuracy, precision, recall of naïve Bayes algorithm(Ilmiah et al., 2018), logistic regression, and support vector machine (Muttaqien, Tibyani, & Hartono, 2022) view in table 11 as follows:

Table 11. Evaluation Accuracy

Algoritma	Sentiment	*Accuracy*	*Precision*	*ROC*
NB	Positive	77,00%	85%	88%
	Negative		42%	64%
	Neutral		100%	100%
LR	Positive	86,85%	87%	85%
	Negative		42%	74%
	Neutral		100%	100%
SVM	Positive	86,42%	86%	87%
	Negative		38%	62%
	Neutral		100%	100%

E. Deployment

After obtaining the data corpus from the evaluation results, the data corpus was entered into a database for the deployment process of the COVID-19 vaccination sentiment analysis application. The following is a flowchart of a sentiment analysis application in figure 3:

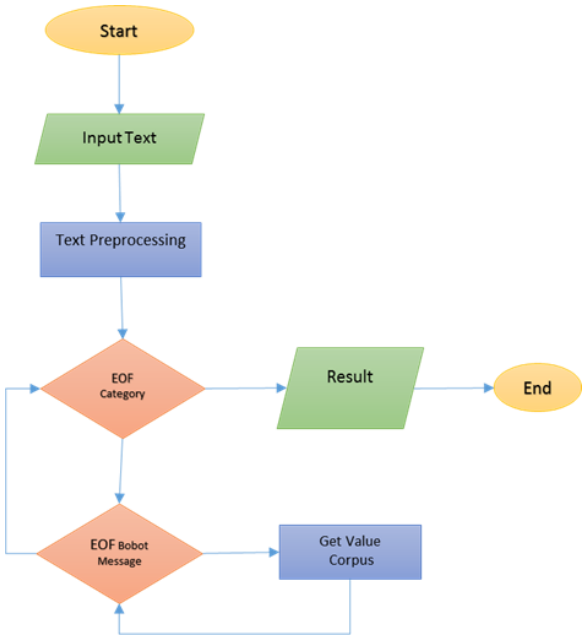

Figure 3. Application Deployment Flowchart

The results of the development of the sentiment analysis application in figure 4 are as follows:

Figure 4. Initial View of Sentiment Analysis Web

In the initial view, the web contains Web Information titled Sentiment Analysis of Covid-19 Vaccination, and the Sentiment Analysis menu is shown in figure 5.

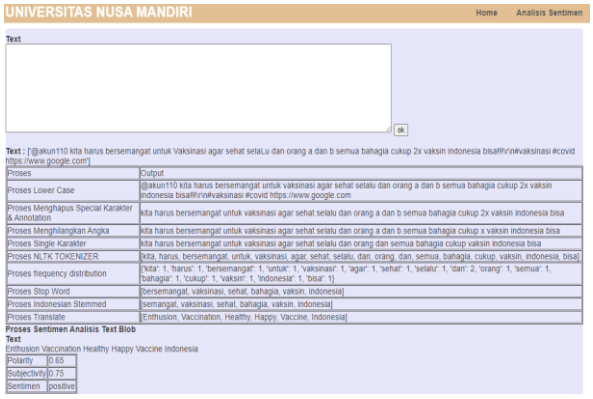

Figure 5. Sentiment Analysis Menu display

The Sentiment Analysis menu has a Text Input menu and an OK button. When we have finished entering the text, the next step is the process of word weighting to produce a conclusion from each word with positive, negative, and neutral sentiments.

CONCLUSION

Based on data obtained from Twitter social media related to the topic of discussion around the Covid-19 vaccination in Indonesian society. The vaccination program held in Indonesia was well received and accepted by the people of Indonesia. The results of sentiment grouping using Textblob are 43% of people with positive sentiments, 40% neutral sentiments, and 16% negative sentiments. The study's results were proven using Naive Bayes, SVM, and Logistic Regression. The highest value for the accuracy of testing with data mining methods is Logistic Regression, which has an accuracy rate of 86.85%. Moreover, from the research results, a sentiment analysis application was built to make it easier to determine sentiment in the community.

REFERENCES

Ahuja, S., & Dubey, G. (2017). Clustering and Sentiment Analysis on Twitter Data. *2017 2nd*

International Conference on Telecommunication and Networks (TEL-NET), *1–5*(1), 1–5. https://doi.org/10.1109/TEL-NET.2017.8343568

Dewi, S. A. E. (2021). Komunikasi Publik Terkait Vaksinasi Covid 19. *Health Care : Jurnal Kesehatan*, *10*(1), 162–167. https://doi.org/10.36763/healthcare.v10i1.119

Ian H. Witten, Eibe Frank, Mark A. Hall, C. J. P. (2019). *Data Mining Practical Machine Learning Tools and Techniques*. (C. Kent, Ed.). Todd Green. Retrieved from https://www.sciencedirect.com/book/9780123748560/data-mining-practical-machine-learning-tools-and-techniques#book-description

Ilmiah, P., Afshoh, F., Informatika, P. S., Komunikasi, F., Informatika, D. A. N., & Surakarta, U. M. (2018). Analisa Sentimen Menggunakan Naïve Bayes. *Jurnal Sains Dan Teknologi*, *10*(2), 2. https://doi.org/https://doi.org/10.32764/saintekbu.v10i2.190

K-means, M. A. (2017). Text Mining Untuk Analisis Sentimen Review Film. *Techno.COM*, *16*(1), 1–8. https://doi.org/10.33633/tc.v16i1.1263

Keahlian, K., & Data, R. (2021). Analisis Sentimen Masyarakat Terhadap COVID-19 Pada Media Sosial, *1*(1), 10–12. https://doi.org/https://doi.org/10.20895/dinda.v1i1.180

Kurniawan, S., Gata, W., Puspitawati, D. A., Parthama, I. K. S., Setiawan, H., S, A., & Hartini. (2019). Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis. *IOP Conference Series: Materials Science and Engineering*, *385*(1), 1. https://doi.org/10.1088/1757-899X/835/1/012057

Makmun, A., & Hazhiyah, S. F. (2020). Kajian Pustaka Tinjauan Terkait Pengembangan Vaksin Covid – 19 Fakultas Kedokteran Universitas Muslim Indonesia. *Molucca Media*, *13*(oktober), 2. https://doi.org/https://doi.org/10.30598/molmed.2020.v13.i2.52

Mart, F., Contreras-ochando, L., & Lachiche, N. (2019). CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories. *IEEE Xplore*, *33*(8), 1. https://doi.org/10.1109/TKDE.2019.2962680

Muttaqien, D. D., Tibyani, T., & Hartono, P. P. (2022). Implementasi Support Vector Machine pada Analisis Sentimen mengenai Bantuan Sosial di Era Pandemi Covid-19 pada Pengguna Twitter. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *10*(1), 6. https://doi.org/http://j - ptiik.ub.ac.id 2548 - 964X

Nur Khormarudin, A. (2016). Teknik Data Mining: Algoritma K-Means Clustering. *Jurnal Ilmu Komputer*, *1*(1), 1–12. Retrieved from https://ilmukomputer.org/category/datamining/

Primadhita, Y., & Budiningsih, S. (2020). Analisis Perkembangan Usaha Mikro Kecil Dan Menengah Dengan Model Vector Auto Regression. *Jurnal Manajemen Kewirausahaan*, *17*(1), 1. https://doi.org/10.33370/jmk.v17i1.396

Rachman, F. F., & Pramana, S. (2020). Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial Twitter, *8*(2), 100–109. https://doi.org/https://doi.org/10.47007/inohim.v8i2.223

Susilo Daniel, P. D. T., & Navarro, J. C. (2021). Performance of Indonesian Ministry of Health in Overcoming Hoax About Vaccination Amid the Covid-19 Pandemic on Social Media. *NYIMAK Journal of Communication*, *5*(1), 1–66. https://doi.org/https://2580-3808

Yolanda, I. (2021). Urgensi Pengaturan Trading In Influence Sebagai Sarana Pembangunan Masyarakat. *DiH: Jurnal Ilmu Hukum*, *6534*(17), 1. https://doi.org/https://doi.org/10.30996/dih.v17i1.4132