Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Sentiment Classification of Fuel Price Rise in Economic Aspects Using Lexicon and SVM Method

Muhammad Fikri Alfauzan[1], Yuliant Sibaroni[2]*, Fitriyani[3]
[1,2,3]Faculty of Informatics Telkom University, Indonesia
[1]alfauzan@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id,
[3]fitriyani@telkomuniversity.ac.id

Submitted : Aug 7, 2023 | **Accepted** : Aug 10, 2023 | **Published** : Oct 1, 2023

Abstract: After being hit by COVID-19 for a long time around the world which resulted in the paralysis of all countries, especially the economic aspects of all countries that dropped dramatically, the world was again shocked by the conflict between Russia and Ukraine which resulted in an increase in world oil prices including in Indonesia, many people complained and opposed the government's policy of increasing fuel prices because fuel affects various aspects, including economic aspects. Based on these problems, researchers use sentiment analysis methods that aim to find out people's opinions on issues that are being discussed throughout Indonesia and this research focuses on comparing the SVM algorithm with TF-IDF feature extraction then using K-Fold Cross Validation after that it is compared with the Lexicon Inset dictionary, in this case the model with Lexicon Inset which contains weighting on each word. In this study, it was found that the dataset model using the SVM algorithm with TF-IDF feature extraction and then using K-Fold Cross Validation obtained an average accuracy of 0.85 using the SVM algorithm. While the model using the automatic labeling dataset using the Indonesian sentiment Lexicon (Lexicon Inset) obtained an average accuracy of 0.68. Classification using SVM with TF-IDF feature extraction is superior to using Lexicon Inset.

Keywords: K-Fold, Lexicon, Sentiment Analysis, SVM, TF-IDF

INTRODUCTION

The existence of conflict causes unstable world oil prices because Russia is one of the world's leading oil producers. The global economic recovery may also be weaker than previously estimated. Post-COVID-19 global economic recovery with the threat of inflation has been seen in several developed countries, from the United States to Indonesia. The increase in fuel prices has the effect of causing an additional 1.9% inflation. With the combination that food is guaranteed to be maintained so that inflation can be kept below 7% until the end of the year, while in developed countries, the increase in energy prices is expected to reduce economic growth by 0.69 proportion points in 2022 and 0.95 percentage points in 2023(Adiah, 2022). To know public's response to the increase in fuel prices, especially on the economic aspect, a method that can be used to find out this kind of public response is sentiment analysis.

Sentiment analysis is a way to assess written or oral opinions to determine whether an opinion is positive, negative or neutral(Anshuman, Rao, & Kakkar, 2017). One of the popular methods or algorithms for classifying sentiment analysis is the Naïve Bayes Algorithm(Mujahidin, Prasetio, & Utomo, 2022), as was done by Saputra in his research on the topic of rising fuel prices(Saputra & Waluyo, 2022). But the SVM Algorithm can also be used for sentiment analysis(Isnain, Sakti, Alita, & Marga, 2021).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Several feature extractions have often been added in sentiment analysis(Buntoro, 2017) by analyzing sentiment regarding the 2017 DKI Jakarta Governor candidate using Lexicon with Naive Bayes and SVM. One of the researchers who used Lexicon extraction features in their research(Sanjaya & Lhaksmana, 2020) then there were researchers who used Lexicon Inset in their research(Shalehanny, Triayudi, & Handayani, 2021).

Punctuation and capital letters in sentences need to be removed by using Cleansing or Case Folding and changing each word in the sentence into a basic word using the Stemming method and knowing the level of accuracy obtained from the classification results using the SVM algorithm, preprocessing data is very important in sentiment analysis because it can improve accuracy(Krouska, Troussas, & Virvou, 2016).

The data used in this study was obtained from Twitter because Twitter is one of the largest social media in the world that rivals Facebook(Basri, 2017), which has now changed its name to X to support the business to be developed by Elon Musk's company(Riyanto, 2023). This research focuses on comparing the performance of SVM classification with TF-IDF feature extraction and the performance of Lexicon Inset classification which has been made a dictionary for each word(Koto & Rahmaningtyas, 2018), from the comparison will be seen the accuracy of the classification performed.

METHOD

The system to be built is a system that can analyze public sentiment on Twitter that has been implemented by Lexicon Inset using TF-IDF feature extraction classified by SVM. The flow of the system built in this study is shown in Figure 1.

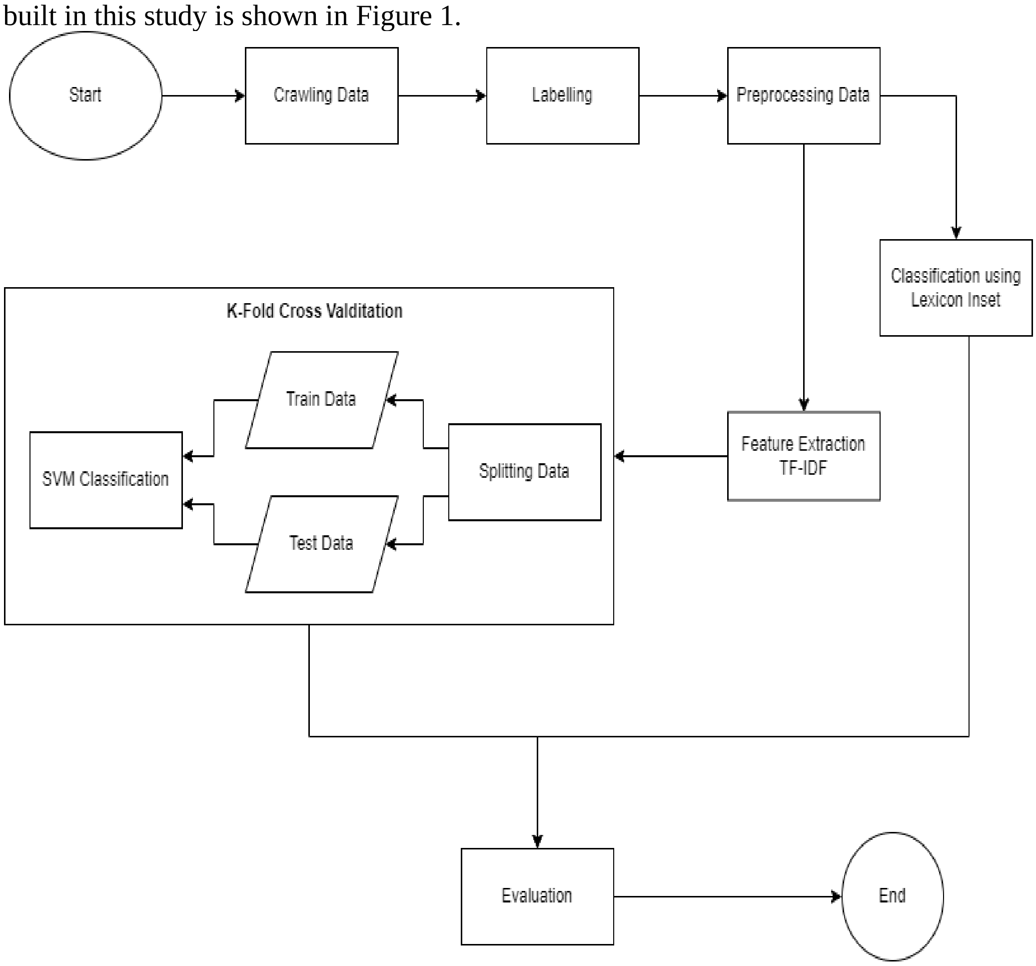

Figure 1 Flowchart of the built system

Crawling Data

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data used from Twitter social media in the time span of September 2022. Data from Twitter is taken using a python application using the snscrape library. The data that was successfully obtained from the crawling process amounted to around 20,000 data, but the dataset used was only 4369. The data collected and used were tweets related to economic aspects, using the keywords economy, poor, rupiah.

Data Labelling

Labelling was done manually by 3 people and labelled positively if the tweet agreed with the government's policy of raising fuel prices and negatively if the tweet disagreed with the government's policy, this is done to produce good accuracy, because quality labelling can provide high accuracy. The labels used are 1 (positive), -1 (negative), the data obtained from this labelling are 1267 positive labelled data, 3102 negative labelled data.

Preprocessing Data

Preprocessing is the stage for processing the dataset into quality data. The goal is to make the data ready for use by eliminating elements that are considered meaningless(Krouska et al., 2016). The preprocessing stages used in this research are case folding, removing punctuation marks, tokenization. Illustration of preprocessing can be seen in Table 1.

Table 1 Illustration Preprocessing

Stage	Description
Case Folding	The process of changing the text into uniform
Remove Punctuation	The process of cleaning documents and selecting unnecessary words such as html, emoticons, hashtags, mentions and urls.
Tokenization	The process of separating or truncation of each word
Stopword	The process removes words that are less effective use stopword indonesia from nltk library
Normalization	The process of converting informal words into formal words by using a dictionary kamus alay(Salsabila, Winatmoko, Septiandri, & Jamal, 2018)
Stemming	The process of cutting words into basic words. At this stage the author uses a sastrawi library

Sentiment Classification using Lexicon Inset

After the data is preprocessed, the data is labeled automatically using the Lexicon Inset by utilizing the frequency of occurrence of the word opinion in each sentence. The features taken are only the words contained in the Lexicon Inset. A positive label is automatically assigned to data with a total score > 0 labeled with 1, while a negative label is automatically assigned to data with a total score < 0 labeled with -1.

Table 2 Illustration of the Classification Process Using Lexicon Inset

word sentences	Lexicon Indonesia Sentiment (Lexicon Inset)	Classify

*name of corresponding author

tolak kenaikan harga bbm	tolak	kenaikan	harga	bbm	Total score	
	-4	0	3	0	-1	Negative
harga bbm mahal merdeka	harga	bbm	mahal	merdeka	Total score	
	3	0	1	0	4	Positive

Feature Extraction TF-IDF

TF is the frequency of occurrence of the term in the dataset and IDF is a weighting of how the term is widely distributed throughout the dataset in question. The smaller the amount of data containing the term in question, the greater the IDF value(Ahuja, Chug, Kohli, Gupta, & Ahuja, 2019). The following is the use of TF-IDF in this study.

$$idf(t, D) = log\left(\frac{N}{df(t)+1}\right) \quad (1)$$

The IDF calculation above is then combined with the TF word weighting so that it becomes the TF-IDF method. The equation used to calculate the TF-IDF weighting of a term is as follows.

$$tf_{ij} = \frac{f_d\,(i)}{f_d\,(j)} \quad (2)$$

Information:
N = Number of all documents
$df(t)$ = Number of documents containing the term
$f_d\,(i)$ = Frequency of occurrence of term i in document j
$max\ f_d\,(j)$= = Total term in document j

Support Vector Machine (SVM) Classification

After doing the weighting at the feature extraction stage, the next step is to build a sentiment classification system using the Support Vector Machine method using TF-IDF for accumulating a word's weight. Feature extraction in classification is expected to be more efficient by identifying features which will then be processed based on the classifier model that has been generated. Furthermore, at this stage requires a set of parameters called hyperparameters. This study implements 3 types of kernels, linear kernels, rbf kernels, and poly kernels. The optimized hyperparameters are C and gamma. At this stage the best score is sought using GridSearchCV and then 1 of the 3 kernels is selected to calculate the accuracy of the SVM method.

Hyperparameter

It is preferable to identify the optimal arguments for the C and gamma parameters before fitting the data to an SVM model, which is best suited for the given problem and this can be done by using grid search. Grid search is a process that allows using the right parameters by trying all the best possible parameters (Deshwal & Sharma, 2019).

SVM Classifier

SVM also has a few kernel functions that is often used which are Linear, Polynomial, Radial Basis

DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Function (RBF). The kernel function is used to change the data into a higher dimension to help the dimension problem (Sunori et al., 2021). Where linear kernel chooses hyperplane and is one of the simple kernels. On the other hand, rbf kernel is used to map the sample in linear number to a higher dimension and a more appropriate multiclass. While polynomial kernel is used to train denormalization (Sunori et al., 2021).

Model Evaluation

At the evaluation stage, system testing is carried out to determine the performance of the classification results that have been made by calculating the accuracy, precision, recall, and f-measure values. Performance evaluation is one of the parameters used to measure how accurate a method is implemented. System development in this evaluation uses a confusion matrix. An overview of the confusion matrix can be seen in Table 3.

Table 3 Confusion Matrix

Predicted Values	Actual Values	
	Positive (1)	Negative (-1)
Positive (1)	TP	FP
Negative (-1)	FN	TN

From Table 3. There are 4 combinations for Predicted Value and Actual Value, namely:
TP is true positive, meaning that the system predicts positive and the actual class value is also positive
TN is true negative, meaning that the system predicts negative and the actual class value is also negative.
FP is a false positive, meaning that the system predicts a positive but the actual class value is negative.
FN is false negative, meaning that the system predicts negative but the actual class value is positive.

These are a few performance metrics that will used, which are:
Accuracy
Accuracy is the ratio of correct predictions from the entire dataset (Ahuja et al., 2019)

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision
Precision is the ratio of correct positive predictions compared to all positive predictive data.

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Recall
Recall is a comparison of True Positive (TP) with the number of actual positive prediction data.

$$recall = \frac{TP}{TP + FN} \tag{5}$$

F1-Score
F1 Score is the harmonic average of precision and recall.

$$f1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{6}$$

RESULT

In this study, the dataset used is a dataset regarding fuel price hikes obtained by using crawling data

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023 e-ISSN : 2541-2019
DOI : https://doi.org/10.33395/sinkron.v8i4.12851 p-ISSN : 2541-044X

on Twitter. After the dataset is collected, the dataset is then cleaned in the preprocessing stage with several stages. The clean dataset will have 2 models, the first model the clean dataset enters the Lexicon Inset stage, then the second enters the TF-IDF feature extraction stage and enters the SVM algorithm using K-Fold. After designing the dataset and training model using SVM, evaluation and comparison are carried out with the Lexicon Inset label.

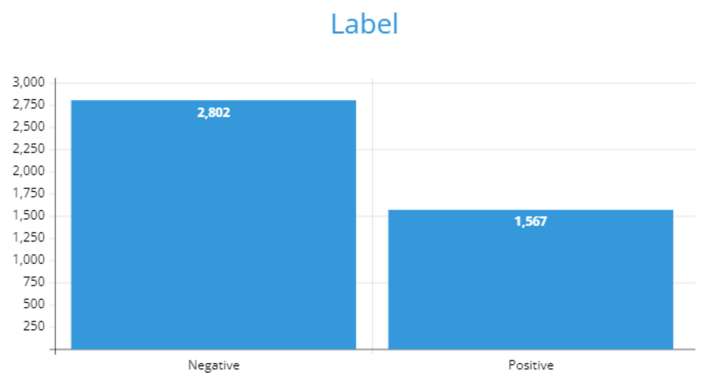

Figure 2 Classification using Lexicon Inset

Figure 2 is the result of data weighting according to the dictionary Lexicon Inset, the dataset obtained from automatic labeling using the Lexicon Inset has 1567 positive and 2802 negative.

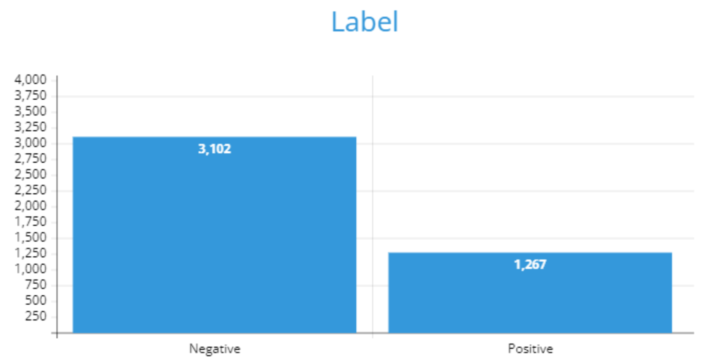

Figure 3 Classification using SVM

In figure 3, the dataset obtained from labelling has 1267 positives and 3102 negatives, this dataset will be used for classification using the SVM algorithm.

Table 4 Confusion Matrix Lexicon Inset

Predicted Values	Actual Values	
	Positive (1)	Negative (-1)
Positive (1)	709	568
Negative (-1)	858	2244

Classification on positive and negative sentiment from the true value and prediction value using confusion matrix. Got the score 709 for true positive, and 2244 for true negatives as shown in Table 4, then the F1 score Lexicon Inset is 0.50 and the accuracy is 0.68.

Best performance model

Table 5 shows the hyperparameter combinations of SVM that will be optimized

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 5 Search Hyperparameter

Parameter	Value
C	0.1
	1
	10
	100
	1000
Kernel	Linear
	Rbf
	Poly
Gamma	1
	0.1
	0.01
	0.001
	0.0001

SVM Testing

Table 6 Linear Kernel

Kernel	C	Gamma	Accuracy
Linear	0.1	1	0.813
		0.1	0.813
		0.01	0.813
		0.001	0.813
		0.0001	0.813
	1	1	0.847
		0.1	0.847
		0.01	0.847
		0.001	0.847
		0.0001	0.847
	10	1	0.792
		0.1	0.792
		0.01	0.792
		0.001	0.792
		0.0001	0.792
	100	1	0.773
		0.1	0.773
		0.01	0.773
		0.001	0.773
		0.0001	0.773
	1000	1	0.545
		0.1	0.545
		0.01	0.545
		0.001	0.545
		0.0001	0.545

In Table 6 by using a linear kernel with several parameters that have been searched for the best accuracy, it can be seen that the gamma parameter does not really affect the accuracy value,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

because the accuracy value in the linear kernel is more influenced by the C parameter.

Table 7 Poly Kernel

Kernel	C	Gamma	Accuracy
Poly	0.1	1	0.725
		0.1	0.710
		0.01	0.710
		0.001	0.710
		0.0001	0.713
	1	1	0.809
		0.1	0.713
		0.01	0.713
		0.001	0.713
		0.0001	0.713
	10	1	0.820
		0.1	0.710
		0.01	0.710
		0.001	0.710
		0.0001	0.710
	100	1	0.819
		0.1	0.725
		0.01	0.710
		0.001	0.710
		0.0001	0.710
	1000	1	0.819
		0.1	0.809
		0.01	0.710
		0.001	0.710
		0.0001	0.710

In Table 7 by using the Poly kernel with several parameters that have been searched for the best accuracy, it can be seen that the poly kernel cannot achieve high accuracy and tends to produce the same accuracy on several parameters and these parameters also do not really affect the accuracy value of the poly kernel.

Table 8 Rbf Kernel

Kernel	C	Gamma	Accuracy
Rbf	0.1	1	0.750
		0.1	0.716
		0.01	0.710
		0.001	0.710
		0.0001	0.710
	1	1	0.849
		0.1	0.832
		0.01	0.723
		0.001	0.710
		0.0001	0.714
	10	1	0.848
		0.1	0.837
		0.01	0.834

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Number 4, October 2023
DOI : https://doi.org/10.33395/sinkron.v8i4.12851
e-ISSN : 2541-2019
p-ISSN : 2541-044X

		0.001	0.725
		0.0001	0.710
	100	1	0.843
		0.1	0.798
		0.01	0.834
		0.001	0.834
		0.0001	0.725
	1000	1	0.843
		0.1	0.798
		0.01	0.786
		0.001	0.833
		0.0001	0.834

In Table 8 by using the rbf kernel with several parameters that have been searched for the best accuracy, it can be seen that the rbf kernel gets varying accuracy values compared to other kernels, it can be seen that the highest accuracy is found in the rbf kernel.

The best parameter obtained is with rbf kernel, C = 1, and gamma = 1, further testing is carried out, in this study the data was tested using 10-fold cross validation, where the data is divided into two, namely 9/10 is training data and the remaining 1/10 is data testing. The test is carried out up to 10 times with successive combinations.

Table 9 Evaluation of Model Accuracy

K	F1-Score	Accuracy
1	0.6667	0.8513
2	0.5506	0.8169
3	0.6912	0.8467
4	0.7426	0.8810
5	0.6562	0.8490
6	0.6774	0.8627
7	0.6064	0.8307
8	0.6032	0.8284
9	0.6772	0.8604
10	0.6804	0.8578

The value obtained classification SVM using TF-IDF resulted in an average F1 score of 0.66, and an average accuracy value of 0.85, this modeling without using Lexicon Inset which only uses SVM algorithm.

In Table 10 we can see a result of the F1 score and the accuracy of SVM and Lexicon

Table 10 Comparison of SVM and Lexicon

	F1-Score	Accuracy
SVM	0.66	0.85
Lexicon Inset	0.50	0.68

DISCUSSIONS

SVM has a variety of parameters that can be used to improve accuracy in testing, these parameters also have their respective functions which if used properly will result in high accuracy, one way that is

*name of corresponding author

often used and quite influential is to use kernels, kernels are used to facilitate data classification by looking for hyperplane that can separate the dataset well, in this study the best parameters were sought from 3 kernels, therefore the authors used hyperparameters to find the best parameters to be used in this SVM test using the GridSearchCV method. As explained above that several parameters have been tried and seen the accuracy value, it can be seen that the highest linear kernel only touches an accuracy of 0.847, while in the poly kernel the highest accuracy is only 0.819, then in the poly kernel touches an accuracy value of 0.849. And the best parameter obtained is the rbf kernel with C = 1 and gamma = 1.

In this study, the results of sentiment classification conducted using the SVM method with TF-IDF feature extraction and tested using the K-Fold Cross Validation technique obtained accuracy results of 0.81 - 0.88 and F1 results of 0.6 - 0.7 with K = 10, this result is the maximum result considering that negative data is much more than positive data.

While classification using Lexicon Inset only gets an accuracy of 0.68 and F1 0.50, the results are quite far compared to classification using SVM, this is because the words in the Lexicon Inset dictionary are limited and not all words have weights in the dictionary.

CONCLUSION

In this study, researchers obtained the results of Lexicon Inset performance and SVM performance in the classification of fuel price increases obtained from Twitter. Experiments show sentiment classification using SVM algorithm with TF-IDF feature extraction that has been searched for the best parameters using hyperparameters, where the best parameter obtained is the rbf kernel whose accuracy is 0.02 higher than the linear kernel. The average accuracy obtained from SVM classification is 0.85 and the F1 value is 0.66 using K-Fold. These results prove that SVM is better than classification using Lexicon Inset which only obtained an accuracy value of 0.68 and an F1 value of 0.50, SVM is superior to Lexicon by a difference of 0.17 for accuracy and 0.16 for F1, this is because the word labeling process is only carried out on words contained in the Lexicon Inset dictionary.

REFERENCES

Adiah, A. (2022). Bank Dunia: Harga Energi Naik, Ekonomi Melambat. Retrieved November 17, 2022, from https://databoks.katadata.co.id/ website: https://databoks.katadata.co.id/datapublish/2022/09/06/bank-dunia-harga-energi-naik-ekonomi-melambat

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, 341–348. https://doi.org/10.1016/j.procs.2019.05.008

Anshuman, Rao, S., & Kakkar, M. (2017). A rating approach based on sentiment analysis. *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, (January), 557–562. https://doi.org/10.1109/CONFLUENCE.2017.7943213

Basri, H. (2017). PERANMEDIA SOSIAL TWITTER DALAM INTERAKSI SOSIAL PELAJAR SEKOLAH MENENGAH PERTAMA DI KOTA PEKANBARU (studi kasus pelajar SMPN 1 kota Pekanbaru). *Strategi Bertahan Hidup Petani Penggarap Di Jorong Sarilamak Nagari Sarilamak Kecamatan Harau Kabupaten Lima Puluh Kota*, *4*(1), 1–13. Retrieved from https://media.neliti.com/media/publications/183768-ID-partisipasi-masyarakat-dalam-pelaksanaan.pdf

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *INTEGER: Journal of Information Technology*, *2*(1), 32–41. https://doi.org/10.31284/j.integer.2017.v2i1.95

Deshwal, V., & Sharma, M. (2019). Breast Cancer Detection using SVM Classifier with Grid Search Technique. *International Journal of Computer Applications*, *178*(31), 18–23. https://doi.org/10.5120/ijca2019919157

Koto, F., & Rahmaningtyas, G. Y. (2018). Inset lexicon: Evaluation of a word list for Indonesian sentiment analysis in microblogs. *Proceedings of the 2017 International Conference on Asian*

*name of corresponding author

Language Processing, IALP 2017, 2018-Janua(December), 391–394. https://doi.org/10.1109/IALP.2017.8300625

Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on Twitter sentiment analysis. *IISA 2016 - 7th International Conference on Information, Intelligence, Systems and Applications*, (November 2017). https://doi.org/10.1109/IISA.2016.7785373

Mujahidin, S., Prasetio, B., & Utomo, M. C. C. (2022). Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM Pada Komentar Youtube Dengan Metode Gaussian naïve bayes. *Voteteknika (Vocational Teknik Elektronika Dan Informatika)*, *10*(3), 17. https://doi.org/10.24036/voteteknika.v10i3.118299

Rahman Isnain, A., Indra Sakti, A., Alita, D., & Satya Marga, N. (2021). Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm. *Jdmsi*, *2*(1), 31–37. Retrieved from https://t.co/NfhnfMjtXw

Salsabila, N. A., Winatmoko, Y. A., Septiandri, A. A., & Jamal, A. (2018). Colloquial Indonesian Lexicon Nikmatun. *Angewandte Chemie International Edition, 6(11), 951–952.*, *3*(1), 10–27. Retrieved from https://medium.com/@arifwicaksanaa/pengertian-use-case-a7e576e1b6bf

Sanjaya, G., & Lhaksmana, K. M. (2020). *Lexicon Based). 7*(3), 9698–9710.

Saputra, R. A., & Waluyo, S. (2022). Penerapan Algoritma Naive Bayes Dalam Analisis Kenaikan Bahan Bakar Minyak Pada Twitter. In *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) Jakarta-Indonesia*. Retrieved from https://senafti.budiluhur.ac.id/index.php

Shalehanny, S., Triayudi, A., & Handayani, E. T. E. (2021). Public'S Sentiment Analysis on Shopee-Food Service Using Lexicon-Based and Support Vector Machine. *Jurnal Riset Informatika*, *4*(1), 1–8. https://doi.org/10.34288/jri.v4i1.287

Sunori, S. K., Singh, D. K., Mittal, A., Maurya, S., Mamodiya, U., & Juneja, P. K. (2021). Rainfall Classification using Support Vector Machine. *Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2021*, 433–437. https://doi.org/10.1109/I-SMAC52330.2021.9640773

*name of corresponding author