Buletin Ilmiah Sarjana Teknik Elektro
Vol. 5, No. 1, March 2023, pp. 132-144
ISSN: 2685-9572, DOI: 10.12928/biste.v5i1.7414

Sentiment Analysis of the Increase in Fuel Prices Using Random Forest Classifier Method

Karandi Nurbagja, Nurirwan Saputra, Ahmad Riyadi, Meilany Nonsi Tentua
Informatics Research Program, Faculty of Science and Technology, Universitas PGRI Yogyakarta, Indonesia

ARTICLE INFORMATION

Article History:

Submitted 20 January 2023
Revised 25 February 2023
Accepted 09 March 2023

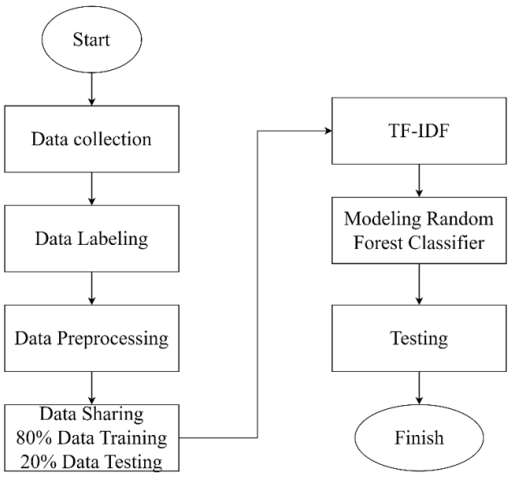
Keywords:

BBM;
Random Forest;
Preprocessing;
Streamlit;
YouTube

Corresponding Author:

Karandi Nurbagja,
Informatics Research Program,
Universitas PGRI Yogyakarta,
Yogyakarta, Indonesia.
Email:
karandinurbagja@gmail.com

ABSTRACT



This research focuses on examining how the economy in Indonesia is affected by the increasing fuel prices, with a particular emphasis on the impact on the lower and middle-income populations. The announcement made by President Jokowi and his ministers about the fuel price hikes has spurred public reactions on social media platforms. The rise in fuel costs has a notable impact on people's livelihoods, especially concerning the surge in prices of essential goods. Sentiment analysis measures public opinion about the increase in fuel prices. The data taken from social media needs to be more balanced. The data is labeled with two identifications, “negative”, “positive” and “neutral”. The method used in sentiment analysis is the random forest method. This research contributes to determining the number of trees used in the model formation with public opinion data. The F1-score measurement result on the model is achieved at a value of 60%.

Document Citation:

K. Nurbagja, N. Saputra, A. Riyadi, and M. N. Tentua, “Sentiment Analysis of the Increase in Fuel Prices Using Random Forest Classifier Method,” *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 5, no. 1, pp. 132-144, 2023, DOI: 10.12928/biste.v5i1.7414.

1. INTRODUCTION

Fuel oil (BBM) is petroleum which is very important and is a basic need for people's lives [1][2]. Fuel oil (BBM) is energy that needs to be subsidized because the price of fuel is strongly influenced by factors external, namely the price of crude oil on the world market [3][4]. Fuel oil (BBM) has a major impact on several sectors of the economy, affecting inflation and changes in prices of basic commodities [5]. Fuel oil (BBM) has a big impact on the people of Indonesia, both for direct and indirect consumption. With this change in fuel prices affecting distribution, transportation, production costs so that it also affects the prices of other goods, especially basic food needs are also affected [6][7]. Rising fuel prices in Indonesia have had a significant impact on the declining economy of the people, especially among the lower middle class [8]. Most people are not aware of the increase in fuel prices which will have gradual implications for the economy in Indonesia which will have an increasingly high impact on poverty will encourage society to give feedback on the increase in fuel prices [9]. The contribution of this research will process public feedback regarding the increase in fuel prices, not only processing positive and negative feedback, but also processing neutral comments. The increase in population has increased which has had various impacts on aspects of life, especially the impact of increasing the need for staple goods [10]. Based on the percentage of poverty from 1990 to 2003 as a result of the monetary crisis shown in Figure 1.

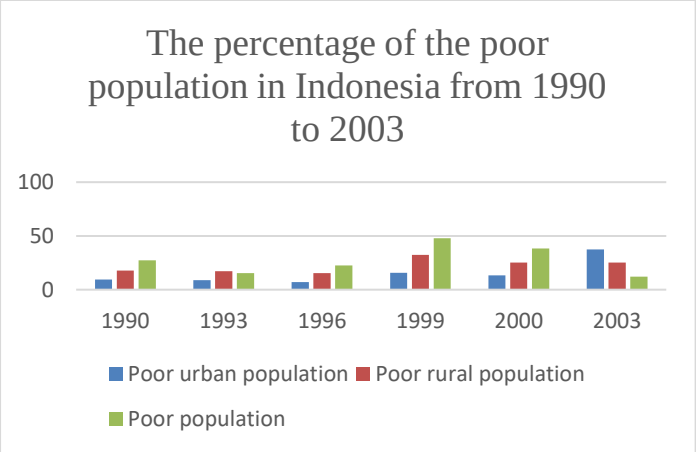


Figure 1. Percentage of the poor population in Indonesia from 1990 to 2003

In the announcement of the increase in fuel prices through President Joko Widodo and Minister Arifin Tasrif who is the Minister of Energy and Mineral Resources of Indonesia announced that the fuel was raised which is shown in Figure 2.

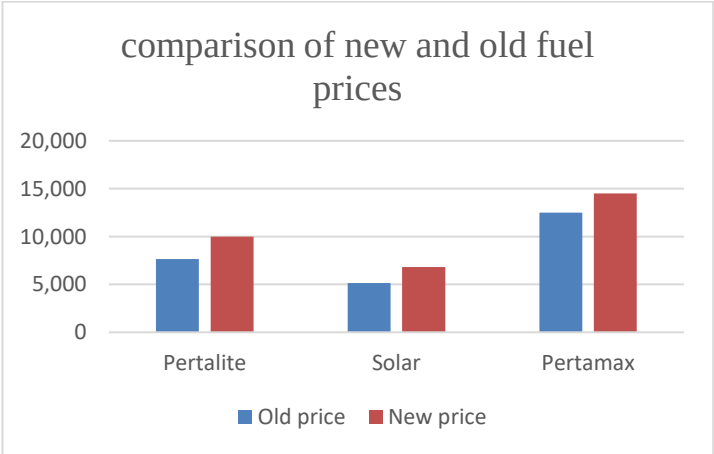


Figure 2. Comparison fuel prices

The increase in fuel that has been announced through social media YouTube. YouTube is one of the applications that is most in demand by the public because it provides a source of information about various topics of issues that are being discussed and also users YouTube can provide comments on information that

is currently being discussed [11]. The topic used in this research is the increase in fuel prices which affects the way the government handles the economic crisis at this time.

Increase in fuel prices through social media YouTube issued by channel CNN Indonesia, entitled "President Jokowi Officially Increases Fuel Prices Today, Pertalite to IDR 10,000, Solar to IDR 6,800" which was uploaded on September 3, 2022 which contains an explanation of the increase in fuel prices which makes people feel miserable and suffer with increase in the price of basic goods affected by the increase in the price of fuel which was conveyed in the comments YouTube and also Due to the policy issued by the government to increase fuel prices, many people are surprised by the increase in fuel prices. Because the government announced a sudden increase in the price of fuel, which caused the public not to be ready for the increase in fuel prices. Regarding the increase in fuel prices, several public opinions through social media YouTube gave several opinions, starting from giving positive opinions to the government but not only positive opinions, there were also those who gave negative opinions because there were deficiencies in government policies in fuel hikes or because of subjectivity. Sentiment analysis is a process that processes textual data automatically which is used to obtain information about fuel price hikes [12][13][14]. Data mining is the science of artificial intelligence, machine learning, and statistics [15]. Obtaining public comments about the increase in fuel prices on social media YouTube used to analyze sentiment on social media by implementing a system so that it can determine Negative, Positive, and Neutral comments on information on rising fuel prices. Community comments will be taken through the process of scraping and cleaning. Scraping by extracting data or information from the intended site in retrieving the dataset on the site or website [16].

The data preprocessing has several steps, such as: remove punctuation, Case Folding, Tokenization, Stemming, and Stopword. Process Preprocessing It is hoped that the data comments that will be processed can produce clean data and be used to see the results of the accuracy of the data to be processed [17]. In this research weighted data using TF-IDF and research methods used for classification uses Random Forest Classifier which is a suitable method for this calculation [18][19]. The contribution of this research will be tested using a confusion matrix to determine value accuracy, precision, recall and f^1score.

2. METHODS

The method that will be used in this research is to take a dataset about "Public comments on the increase in fuel prices in Indonesia" that predicts the accuracy of netizens' responses using the Random Forest Classifier method is conducted through several stages, starting with data collection through scraping process. After the data collection, the labeling process is performed, which provides the values of positive, negative, and neutral comments. After the labeling process that assigns a class to each comment, the preprocessing stage is conducted, which includes several steps: removal of punctuation, lowercase conversion, tokenization, stemming, and stopword removal. After the pre-processing stage, the data is weighted using TF-IDF and processed into a Random Forest Classifier model. Finally, the processed data is tested as shown in Figure 3.

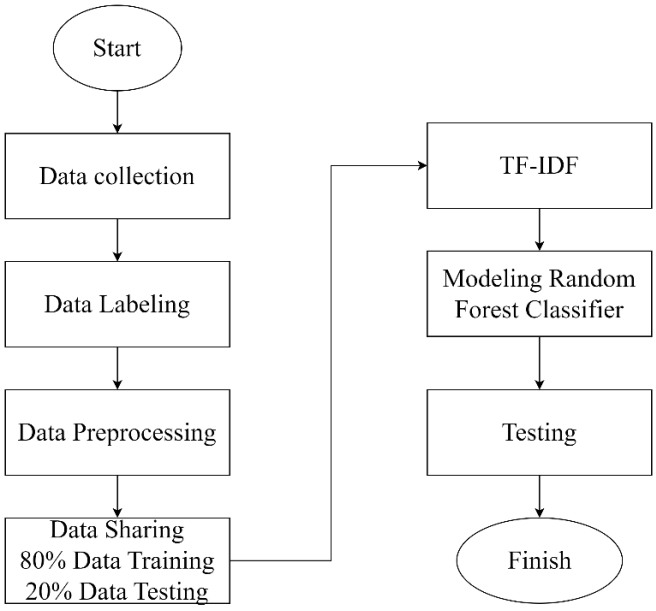


Figure 3. Flowchart methodology

2.1. Data Collection

This research first involves collecting data from YouTube by scraping comments without using the YouTube API. The dataset is related to "CNN Indonesia News on the Increase in Fuel Prices" and concerns public opinion on YouTube regarding the explanation of the increase in fuel prices by President Joko Widodo. The research samples a random dataset of 1804 responses from netizens obtained from YouTube. The news was uploaded on September 3, 2022 and contains the most comments until September 2022.

2.2. Data Labeling

This process assigns labels to words such as Negative, Neutral, and Positive, in order to calculate the accuracy of the sentiment obtained automatically. This process will be done manually [20]. The labels given to the dataset used include:

- Neutral words are given the designation of number (0): Comments that contain opinions that are not related to the theme or topic of discussion in the research taken by the researcher.
- Negative words are given the designation of number (1): Comments that give a negative response to the policy of increasing fuel prices and can have a negative impact on readers or society, which can trigger good opinions about the policy.
- Positive words are given the designation of number (2): Comments that contain public approval of the increase in fuel prices and support the government policy that was made.

In the labeling process itself, it is important to pay attention to the labeling guidelines, in order to avoid errors in assigning labels to each comment data displayed in Table 1.

Table 1. Guidelines labeling

Label	Explanation
Positive	Comments that contain public approval of the increase in fuel prices and support for government policies
Negative	Comments that provide negative responses to the policy of increasing fuel prices and can have a negative impact on readers or the public, which may trigger unfavorable opinions towards the policy
Neutral	Comments that contain opinions that are not relevant to the theme or topic

2.3. Data Preprocessing

In this research, a preprocessing step will be performed to clean the sentences by following several stages, as follows:

***Remove of punctuation*:** Remove of punctuation is the process of removing punctuation, numbers, and characters in words that are not important [19]. This process will make it easier in the next preprocessing stage [21].

***Lowercase*:** Lowercase done in process case folding, lowercase is the change of capital letters to lowercase [13]. The uni process is carried out after the calculation TF-IDF, the use of this process aims to change the entire sentence into lowercase letters [20]. The use of this process also aims to increase the level of accuracy of the data retrieved and make alignment of each sentence.

Tokenization: Tokenization is a truncation process of string input or sentences on the data into syllables with the specified number [22][23]. On this process tokenization use unigram token, the purpose is used unigram token because it will be easier to carry out the next process or the word weighting process and improve better accuracy [24]. Unigram token is token which separates the sentence into one syllable. With unigram token the vocabulary used will be more visible in value [25].

Stemming: Stemming is the process of reducing words to their base form by removing suffixes in order to reduce the number of words and improve the accuracy of the analysis. In Indonesian language, it is used to separate words into their appropriate verb form in the Great Indonesian Dictionary (KBBI) [1]. This process will use the Sastrawy Master library [5] which is a simple python library that provides Indonesian word stemming [26]. The purpose of stemming is to remove inflections from words. This will greatly affect the calculation of the number of words that will be done during TF-IDF [27].

Stopword removal: This process will reduce the number of words that will not be used in the data [27][28]. This step is performed after tokenization will determine the words that are not included in the connecting words used in the sastrawi master by removing the connecting words or inflections found in the sastrawi master library with important words from the token results are taken using a stoplist algorithm (removing less important words) or wordlist (storing important words) [29] as shown in Figure 4.

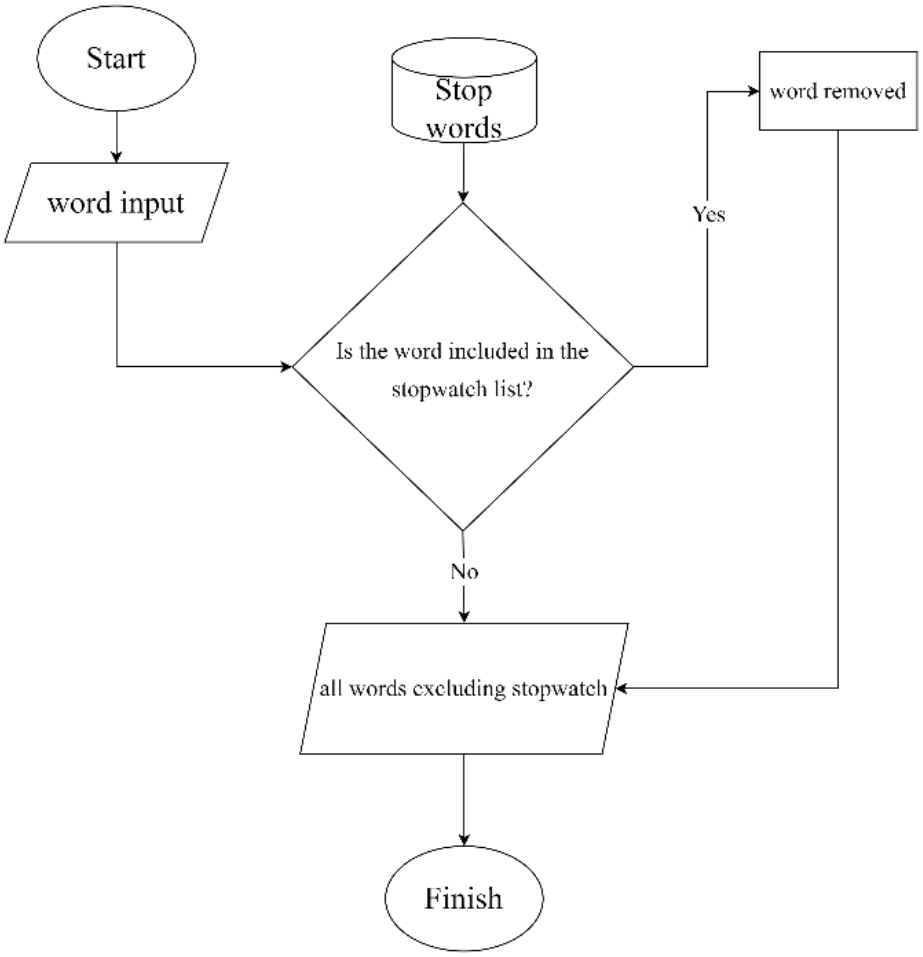


Figure 4. Stopword removal flowchart

2.4. Classification by Method Random Forest Classifier

This research performs classification using the Random Forest Classifier method to determine the accuracy, precision, recall, and F1 score. Random Forest is a method developed from the CART (Classification and Regression Trees) method, which is also a method or algorithm of decision tree techniques [29][30][31]. Random Forest is a collection of several trees, where each tree depends on the pixel value of each vector taken at random and independently. This method can process quickly and allows to process trees as much as desired [32]. This research will use testing with confusion matrix validation.

The prediction model that uses the Random Forest Classifier method is carried out by searching for comments from the dataset that has undergone scraping and preprocessing processes. The formula used to predict in the Random Forest Classifier method uses entropy and information gain [33][34]. Here is the entropy formula (1).

$$Entropi(S) = E(S) = \sum_{j=1}^{k} -p_j log_2 p_j \quad (1)$$

S is a case set, n is number of partitions S, Pi is proportion of Si to S.

The formula for information gain is (2).

$$Gain(A) = Entropi(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \; x \; Entropi(S_i) \quad (2)$$

S is a case set, A is the feature, n is number of partitions atribute A, $|Si|$ is proportion of Si to S, $|S|$ is number of cases S.

The calculations above will be done to find the gain on each data taken using bagging method [35]. Bagging method it self is a random data sampling method which allows the data to be selected again in the next data calculation process.

In this research, two attributes will be calculated manually using the entropy and information gain formulas. If one attribute has the highest gain value, that attribute will become a node or tree. Then the remaining attributes will be calculated further by adding another attribute. When one attribute results in a value of "1", the data will be recalculated. On the other hand, if the classification result is "0", no further calculation is necessary. This process is repeated during the research to produce a total of 50 trees from the Random Forest Classifier classification method, then a voting process will be carried out to determine the number of labels formed from several of those trees.

In classification using the Random Forest Classifier method, the purpose of the algorithm is to build a classifier with the given labels, specifically in this research, a total of 3 labels will be processed from preprocessing, modeling, and testing. In the preprocessing process, it is done by removing punctuation, lowercasing, tokenizing (unigram), stemming, and stopword removal. Then, the Random Forest Classifier method will be modeled and the data will be divided into 80% training data and 20% testing data.

The Random Forest Classifier classification is tested using the multi-class classification confusion matrix validation method. To calculate accuracy, the formula used is (3).

$$Accuracy = \frac{TP + TN + TNET}{TP + TN + FP + FN + TNET + FNET} \quad (3)$$

To calculate precision, the formula used is (4).

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

To calculate recall, the formula used is (5).

$$Recall = \frac{TP}{TP + FN + FNET} \quad (5)$$

F^1score is a measure of a model's accuracy that takes both precision and recall into account. It is calculated using the following formula (6).

$$F^1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Note Information shown in Table 2.

Table 2. Note formula confusion matrix

Symbol	Description
TP:	The results of a positive prediction and its actual results are also positive
TN:	The results of a negative prediction and its actual results are also negative
TNET:	The results of a neutral prediction and its actual results are also neutral
FP:	The results of a positive prediction but its actual results are negative or neutral
FN:	The results of a negative prediction but its actual results are positive or neutral
FNET:	Theresults of a neutral prediction but its actual results are positive or negative

True Positif (TP) is the results of a positive prediction and its actual results are also positive. **True Negatif (TN)** is the results of a negative prediction and its actual results are also negative. **True Netral (TNET)** is the results of a neutral prediction and its actual results are also neutral. **False Positif (FP)** is the results of a positive prediction but its actual results are negative or neutral. **False Negatif (FN)** is the results of a negative prediction but its actual results are positive or neutral. **False Netral (FNET)** is the results of a neutral prediction but its actual results are positive or negative

3. RESULT AND DISCUSSION

3.1. Data Labeling

With a lot of data taken as many as 1804 data comments that go through a labeling process are given guidelines or labels containing numbers that will make labeling easier. This figure is a label containing negative information for number 1, neutral for number 0 and positive for number 2. This resulted in 907 negative sentiments, 402 neutral sentiments and 402 positive sentiments regarding the increase in fuel prices. Process Cleaning used for cleaning unnecessary words or punctuation from comment data taken from

YouTube. Then the results of labeling the data are shown in Table 3. This research contains comments that are negative, neutral and positive as shown in Figure 5.

Table 3. Data labeling

No	Comments	Label
1	Saya rasa cuman kamu dan seglintir komplotan mu yg menderita	Negative
2	nah itu masalahnya...terus perbaiki regulasi dan pengawasan bukan malah naikin hargax,,,,kecuali kl alasannya negara bangkrut ya apa blh buat kita tetap ikhlas BBM naik	Positive
3	Berdoa dan berusaha gimana caranya usaha gak bangkrut	Neutral

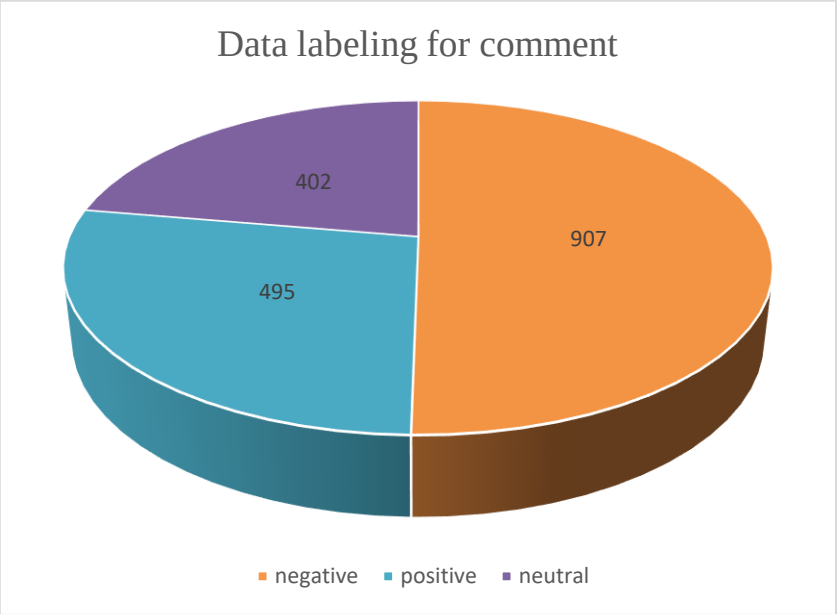


Figure 5. Graph for data labeling for comment

3.2. Data Preprocessing

Process Cleaning used for cleaning unnecessary words or punctuation from comment data taken from YouTube. The process above will produce data in Table 4.

Table 4. Example data prepocessing

Stages	Results
Before Preprocessing	Semangaaat guys, semoga seluruh rakyat Indonesia sehat2 selalu dan lancar rezekinya
Remove of Punctuation	Semangaaat guys semoga seluruh rakyat Indonesia sehat selalu dan lancar rezekinya
Lowercase	semangaaat guys semoga seluruh rakyat indonesia sehat selalu dan lancar rezekinya
Tokenization (Unigram)	'semangaaat', 'guys', 'semoga', 'seluruh', 'rakyat', 'indonesia', 'sehat', 'selalu', 'dan', 'lancar', 'rezekinya'
Stemming	'semangaaat', 'guys', 'moga', 'seluruh', 'rakyat', 'indonesia', 'sehat', 'selalu', 'dan', 'lancar', 'rezeki'
Stopword	'semangaaat', 'guys', 'moga', 'rakyat', 'indonesia', 'sehat', 'lancar', 'rezeki'

3.3. Analysis Sentiment Model

The data used in this study amounted 1807. Manual calculation using several data from the data training an then it will be tested using the testing data and resulting in 6 trees as the reserach output, as shown in Table 5 and Figure 6.

Table 5. Result of Tree

Data	Tree
Training	6

	A	B	C	D	E	F
1	K1	K2	K3	K4	K5	Label
2	turun	harga	bbm	jangan	naik	negatif
3	bbm	naik	bantu	hutang	negara	positif
4	gaji	dpr	gede	banget	guys	positif
5	turun	harga	bbm	lonjak	naik	negatif
6	bagus	subsidi	bbm	di	cabut	negatif
7	ngga	sengsara	kalau	kerja	keras	positif
8	emang	sengsara	bbm	tiap	hari	netral
9	gaji	dikit	di	syukuri	aja	positif
10	hitung	sendiri	aja	bang	selisih	netral
11	rakyat	sengsara	bbm	naik	terus	negatif
12						

Figure 6. Data training

The research will take example of the tree generated from this study, as shown in Figure 7, Figure 8, Figure 9, and Figure 10.

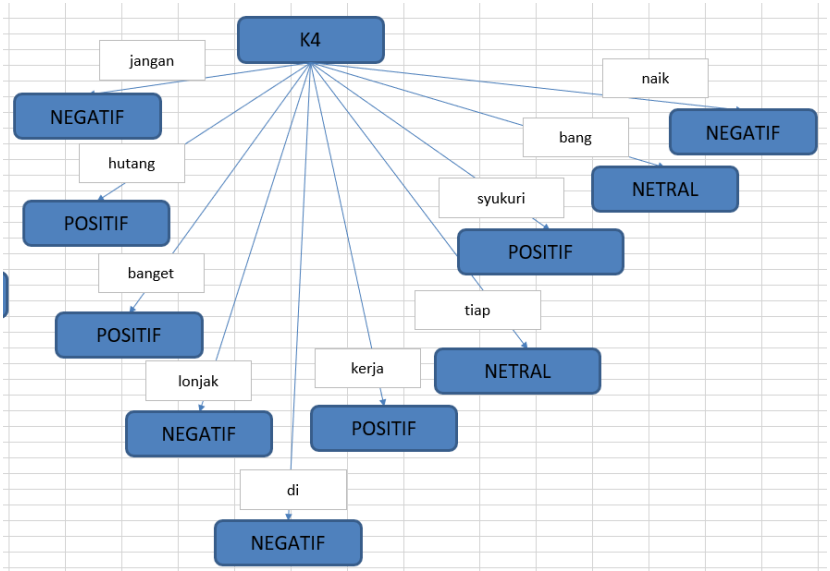


Figure 7. Example of tree 1

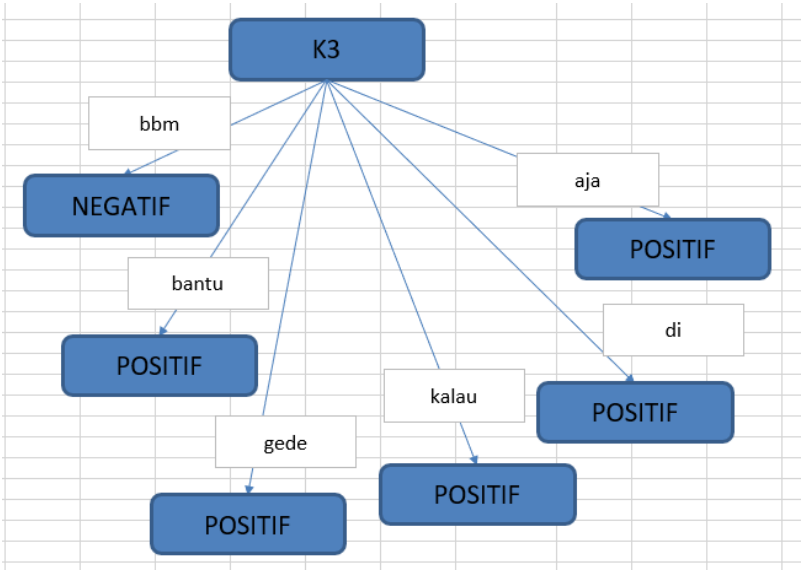


Figure 8. Example of tree 2

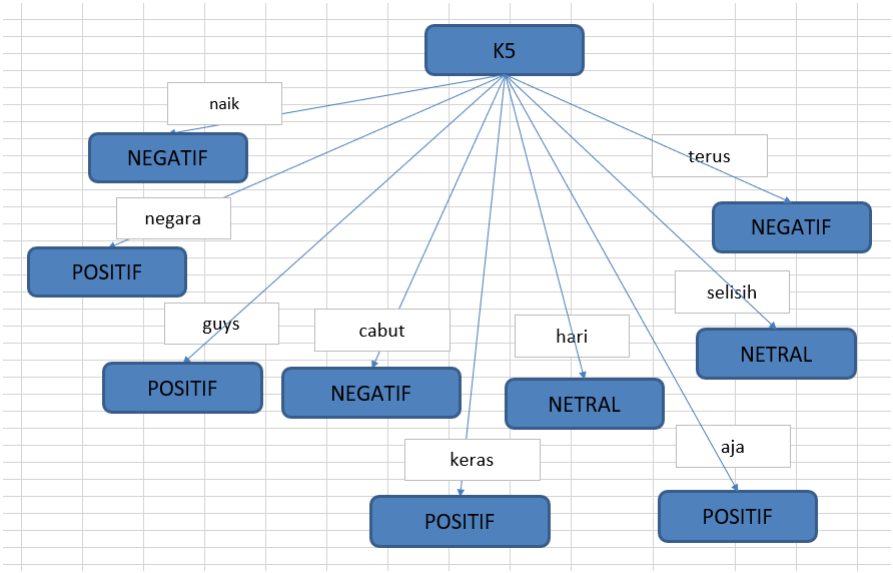


Figure 9. Example of tree 3

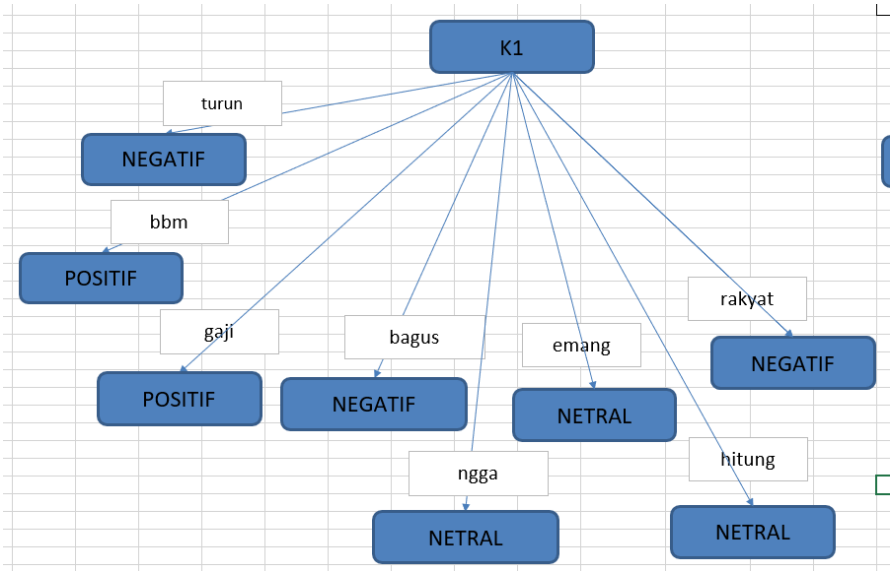


Figure 10. Example of tree 4

After the tree is formed, the testing data is ready to be examined by checking each word in the decision tree output shown in Table 6.

Table 6. Try to testing

K1	K2	K3	K4	K5
Turun	Harga	BBM	Lonjak	Naik

Then each word will be examined according to the result obtained from the decision tree, and a voting process will be conducted to see the most frequent classification. In that sentence, there are 4 negative words shown in Table 7.

Table 7. Adjustment with the decision tree

Negatif	4
Positif	0
Netral	0

This, is can be concluded that the sentence formed by the the testing data is classified as negative based on the decision tree output shown in Table 8.

Table 8. The result of the testing data

K1	K2	K3	K4	K5	Label
Turun	Harga	BBM	Lonjak	Naik	Negatif

3.4. Testing

After testing, it will produce the accuracy, precision, recall, and F1-score values of the confusion matrix shown in Table 9.

Table 9. Result confusion matrix

	NET	NEG	POS
NET	44	28	11
NEG	22	135	14
POS	13	56	36

Note Table 10, Neg is a designation Negative, Net is a Neutral, and Pos is a Positive.

The percentage of accuracy produced by the Random Forest Classifier method is 60%. The percentage of precision, recall, and f^1 score for each label can be seen in Table 10. This result indicates that the classifier is able to correctly classify 60% of the data based on the labels provided in the research. The precision, recall, and F^1 score values in Table 10 give more detailed information about the classifier's performance for each label. Precision measures the proportion of true positive predictions among all positive predictions made by the classifier. Recall measures the proportion of true positive predictions among all actual positive instances. F1 score is a measure that combines precision and recall, it is the harmonic mean of precision and recall.

Table 10. Percentage result of each label

Label	Random Forest Classifier		
	Accuracy		
	60%		
	Precision	**Recall**	**F^1score**
Positive	59%	34%	43%
Negative	62%	79%	69%
Neutral	56%	53%	54%

The predictive model of this research chose Unigram tokenization with an accuracy of 60% and the largest percentage value obtained for the negative label, which is precision 62%, recall 79%, and f1 score 69%.

The accuracy result is 60%, while the Random Forest Classifier method obtained an accuracy of 74%, because the research used 3 labeling, namely positive, negative, and neutral, while the Random Forest Classifier only used 2 labeling, positive and negative. This research was also conducted using multi-class classification testing or 3x3 confusion matrix testing, because it used 3 labels, while in the Random Forest Classifier method only 2 labels were used.

4. CONCLUSIONS

In conclusion, this research used the Random Forest Classifier method to classify comments on the CNN Indonesia News dataset regarding the increase in fuel prices in Indonesia. The research collected data through scraping and preprocessing 1792 comments. The most common classification was negative. The research found that the Random Forest Classifier method had good accuracy, precision, recall, and F1 score values. The overall accuracy of the classifier was 60%, with the highest percentage value obtained for the negative label, which is precision 62%, recall 79%, and f1 score 69%. This research shows that the use of preprocessing steps and the Random Forest Classifier method is effective in classifying comments on the increase in fuel prices in Indonesia.

5. FUTURE WORKS

This research future works several recommendations for future studies. This research hopes to implement automatic scraping that collects public comments on government policies and processes them with several preprocessing stages. This research is expected to be improved with better methodologies to achieve a higher accuracy level. This research is also expected to be continued with a comparison of other related sentiment analysis methods.

REFERENCES

[1] N. T. S. Saptadi, A. Suyuti, A. A. Ilham and I. Nurtanio, "Energy Potential Estimation System Model To Produce Alternative Energy Briquettes," *2022 International Conference on Informatics Electrical and Electronics (ICIEE)*, pp. 1-7, 2022, https://doi.org/10.1109/ICIEE55596.2022.10009988.

[2] M. Herlina, "Panel Cointegration Analysis in Determining Relationship of Agricultural Commodity and Oil Fuel Price in Indonesia," *Indonesian Journal of Statistics and Its Applications*, vol. 4, no. 2, pp. 341–358, 2020, https://doi.org/10.29244/ijsa.v4i2.662.

[3] S. Sultan, J. J. Sarungu, A. M. Soesilo and S. A. T. Rahayu, "Oil price and Indonesian economic growth," *Problems and Perspectives in Management*, vol. 17, no. 1, pp. 152-162, 2019, https://doi.org/10.21511/ppm.17(1).2019.14.

[4] P. A. Rakhmanto, "Simulation of the Impacts of Fuel Pricing Policies towards Inflation in Indonesia," *International Journal of Multidisciplinary and Current Educational Research (IJMCER)*, vol. 4, no. 4, pp. 24–31, 2022, https://www.ijmcer.com/volume-4-issue-4/.

[5] E. Prabowo, H. Harianto, B. Juanda and D. Indrawan, "The Economic Price of Liquid Petroleum Gas, Poverty and Subsidy Removal Compensation Scenario in Indonesia," *International Journal of Energy Economics and Policy*, vol. 12, no. 5, pp. 169–177, 2022, https://doi.org/10.32479/ijeep.13356.

[6] T. Wang and B. Lin, "Fuel consumption in road transport: A comparative study of China and OECD countries," *Journal of Cleaner Production*, vol. 206, pp. 156-170, 2019, https://doi.org/10.1016/j.jclepro.2018.09.092.

[7] A. Akhmad and A. Amir, "Research of fuel oil supply and consumption in indonesia," *International Journal of Energy Economics and Policy*, vol. 8, no. 4, pp. 13–20, 2018, https://econjournals.com/index.php/ijeep/article/view/6448.

[8] B. Kharisma, "Can A School Operational Assistance Fund Program (BOS) Reduce School Drop-Outs During The Post-Rising Fuel Prices In Indonesia ? Evidence From Indonesia," *Munich Personal RePEc Archive*, no. 70041, pp. 1–14, 2016, https://mpra.ub.uni-muenchen.de/70041/.

[9] A. N. Kurniawan, A. I. Rifai and M. Rizal. S, "Phenomena of Transportation to Work Mode Choice , Due to The Increase of Oil Prices in Indonesia : A Case Light Rail Transit Depot Project Office-Jakarta," *CITIZEN:JurnalIlmiahMultidisiplinIndonesia*, vol. 2, no. 5, pp. 785–793, 2022, https://journal.das-institute.com/index.php/citizen-journal/article/view/193.

[10] A. Akhmad, B. Romadhoni, K. Karim, M. J. Tajibu and M. Syukur, "The Impact of Fuel Oil Price Fluctuations on Indonesia's Macro Economic Condition," *International Journal of Energy Economics and Policy*, vol. 9, no. 2, pp. 277–282, 2019, https://doi.org/10.32479/ijeep.7470.

[11] R. A. Maisal, A. N. Hidayanto, N. F. Ayuning Budi, Z. Abidin and A. Purbasari, "Analysis of Sentiments on Indonesian YouTube Video Comments: Case Study of The Indonesian Government's Plan to Move the Capital City," *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pp. 121-124, 2019, https://doi.org/10.1109/ICIMCIS48181.2019.8985228.

[12] Y. Pratama, A. R. Tampubolon, L. D. Sianturi, R. D. Manalu and D. F. Pangaribuan, "Implementation of sentiment analysis on Twitter using Naïve Bayes algorithm to know the people responses to debate of DKI Jakarta governor election," In *Journal of Physics: Conference Series*, vol. 1175, no. 1, p. 012102, 2019, https://doi.org/10.1088/1742-6596/1175/1/012102.

[13] J. Asian, M. Dholah Rosita and T. Mantoro, "Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods," *JOIN (Jurnal Online Informatika)*, vol. 7, no. 1, p. 132, 2022, https://doi.org/10.15575/join.v7i1.900.

[14] Y. Al Amrani, M. Lazaar and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science*, vol. 127, pp. 511–520, 2018, https://doi.org/10.1016/j.procs.2018.01.150.

[15] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Social Science Research*, vol. 110, p. 102817, 2023, https://doi.org/10.1016/j.ssresearch.2022.102817.

[16] P. Kaur, "Sentiment analysis using web scraping for live news data with machine learning algorithms," *Materials Today: Proceedings*, vol. 65, pp. 3333-3341, 2022, https://doi.org/10.1016/j.matpr.2022.05.409.

[17] A. P. Widyassari, E. Noersasongko, A. Syukur and Affandy, "The 7-Phases Preprocessing Based On Extractive Text Summarization," *2022 Seventh International Conference on Informatics and Computing (ICIC)*, pp. 1-8, 2022, https://doi.org/10.1109/ICIC56845.2022.10006998.

[18] N. O. F. Daeli and A. Adiwijaya, "Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor," *Journal of Data Science and Its Applications*, vol. 3, no. 1, pp. 1–7, 2020, https://doi.org/10.34818/jdsa.2020.3.22.

[19] S. Khomsah, R. D. Ramadhani and S. Wijaya, "The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 6, no. 3, pp. 352–358, 2022, https://doi.org/10.29207/resti.v6i3.3711.

[20] N. Saputra, K. Nurbagja and T. Turiyan, "Sentiment Analysis of Presidential Candidates Anies Baswedan and Ganjar Pranowo Using Naïve Bayes Method," *Jurnal Sisfotek Global*, vol. 12, no. 2, pp. 114-119, 2022, http://dx.doi.org/10.38101/sisfotek.v12i2.552.

[21] F. Y. A'la, "Indonesian Sentiment Analysis towards MyPertamina Application Reviews by Utilizing Machine Learning Algorithms," *Journal of Informatics Information System Software Engineering and Applications (INISTA)* vol. 5, no. 1, pp. 80–91, 2022, https://journal.ittelkom-pwt.ac.id/index.php/inista/article/view/838.

[22] C. Ding *et al.*, "Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 1, pp. 1-34, 2019, https://doi.org/10.1145/3325885.
[23] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp. 46–50, 2018, http://doi.org/10.11591/ijeecs.v12.i1.pp46-50.
[24] A. Ridwan, H. H. Nuha and R. Dharayani, "Sentiment Analysis of Floods on Twitter Social Media Using the Naive Bayes Classifier Method with the N-Gram Feature," *2022 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 114-118, 2022, https://doi.org/10.1109/ICoDSA55874.2022.9862827.
[25] S. Taj, B. B. Shaikh and A. Fatemah Meghji, "Sentiment Analysis of News Articles: A Lexicon based Approach," *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1-5, 2019, https://doi.org/10.1109/ICOMET.2019.8673428.
[26] S. Pebiana *et al.*, "Experimentation Of Various Preprocessing Pipelines For Sentiment Analysis On Twitter Data About New Indonesia's Capital City Using SVM And CNN," *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pp. 1-6, 2022, https://doi.org/10.1109/O-COCOSDA202257103.2022.9997982.
[27] J. Singh and V. Gupta, "A novel unsupervised corpus-based stemming technique using lexicon and corpus statistics," *Knowledge-Based Systems*, vol. 180, pp. 147-162, 2019, https://doi.org/10.1016/j.knosys.2019.05.025.
[28] M. Guia, R. R. Silva and J. Bernardino, "Comparison of Naive Bayes, support vector machine, decision trees and random forest on sentiment analysis," In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, pp. 525–531, 2019, https://doi.org/10.5220/0008364105250531.
[29] A. Y. Clara, A. Adiwijaya and M. D. Purbolaksono, "Aspect Based Sentiment Analysis on Beauty Product Review Using Random Forest," *Journal of Data Science and Its Applications*, vol. 3, no. 2, pp. 67–77, 2020, https://doi.org/10.34818/jdsa.2020.3.58.
[30] M. M. Chen and M. C. Chen, "Modeling road accident severity with comparisons of logistic regression, decision tree and random forest," *Information*, vol. 11, no. 5, p. 270, 2020, https://doi.org/10.3390/info11050270.
[31] A. Nayak and S. Natarajan, "Comparative research of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter Feeds," *International Journal of Advanced Studies in Computer Science and Engineering*, vol. 5, no. 1, pp. 14–17, 2016, http://www.ijascse.org/volume-5-theme-based-issue-1/Support_vector_machine.pdf.
[32] V. Jain and A. Phophalia, "M-ary Random Forest-A new multidimensional partitioning approach to Random Forest," *Multimedia Tools and Applications*, vol. 80, pp. 35217-35238, 2021, https://doi.org/10.1007/s11042-020-10047-9.
[33] M. M. Kholil, F. Alzami and M. A. Soeleman, "AdaBoost Based C4.5 Accuracy Improvement on Credit Customer Classification," *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 351-356, 2022, https://doi.org/10.1109/iSemantic55962.2022.9920463.
[34] M. N. Tentua and A. Sihabuddin, "Improved C4. 5 Algorithm Using the L'Hospital Rule and Prunning on the Recommendation System," vol. 8, no. 11, 2019, http://www.ijstr.org/final-print/nov2019/Improved-C45-Algorithm-Using-The-Lhospital-Rule-And-Prunning-On-The-Recommendation-System.pdf.
[35] N. Bahrawi, "Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based," *Journal of Information Technology and Its Utilization*, vol. 2, no. 2, pp. 29-33, 2019, https://doi.org/10.30818/jitu.2.2.2695.

AUTHOR BIOGRAPHY

Karandi Nurbagja currently completing undergraduate education in the informatics study program at Universitas PGRI Yogyakarta in 2023.

Nurirwan Saputra completed undergraduate education in the informatics engineering study program at the Universitas Islam Indonesia and completed masters education in the electrical engineering study program at Universitas Gajah Mada. Currently author 2 is a permanent lecturer in the informatics study program at the Universitas PGRI Yogyakarta. His research area is sentiment analysis.

Ahmad Riyadi completed undergraduate education in the mathematics study program at Universitas Gajah Mada and completed masters education in the computer science study program at Universitas Gajah Mada. Currently author 3 is a permanent lecturer in the informatics study program at Universitas PGRI Yogyakarta. His research area is computer science.

Meilany Nonsi Tentua completed his undergraduate education at Universitas Gajah Mada mathematics study program and completed his master's degree at Universitas Gajah Mada electrical engineering study program. Currently author 2 is a permanent lecturer in the informatics study program at the Universitas PGRI Yogyakarta. His research areas are Artificial intelligence, machine learning, Data science and Natural Language Processing.