http://doi.org/10.35784/iapgos.3740 *received: 13.06.2023 | revised: 20.08.2023 | accepted: 25.09.2023 | available online: 30.09.2023*

UNBALANCED MULTICLASS CLASSIFICATION WITH ADAPTIVE SYNTHETIC MULTINOMIAL NAIVE BAYES APPROACH

Fatkhurokhman Fauzi, Ismatullah, Indah Manfaati Nur
Universitas Muhammadiyah Semarang, Department of Statistics, Semarang, Indonesia

Abstract. *Opinions related to rising fuel prices need to be seen and analysed. Public opinion is closely related to public policy in Indonesia in the future. Twitter is one of the media that people use to convey their opinions. This study uses sentiment analysis to look at this phenomenon. Sentiment is divided into three categories: positive, neutral, and negative. The methods used in this research are Adaptive Synthetic Multinomial Naive Bayes, Adaptive Synthetic k-nearest neighbours, and Adaptive Synthetic Random Forest. The Adaptive Synthetic method is used to handle unbalanced data. The data used in this study are public arguments per province in Indonesia. The results obtained in this study are negative sentiments that dominate all provinces in Indonesia. There is a relationship between negative sentiment and the level of education, internet use, and the human development index. Adaptive Synthetic Multinomial Naive Bayes performed better than other methods, with an accuracy of 0.882. The highest accuracy of the Adaptive Synthetic Multinomial Naive Bayes method is 0.990 in Papua Barat Province.*

Keywords: adaptive synthetic, classification, imbalance data, accuracy

NIEZRÓWNOWAŻONA KLASYFIKACJA WIELOKLASOWA Z ADAPTACYJNYM SYNTETYCZNYM WIELOMIANOWYM NAIWNYM PODEJŚCIEM BAYESA

Streszczenie. *Należy przyjrzeć się i przeanalizować opinie związane z rosnącymi cenami paliw. Opinia publiczna jest ściśle związana z polityką publiczną Indonezji w przyszłości. Twitter jest jednym z mediów, których ludzie używają do przekazywania swoich opinii. Niniejsze badanie wykorzystuje analizę nastrojów, aby przyjrzeć się temu zjawisku. Opinia jest podzielona na trzy kategorie: pozytywną, neutralną i negatywną. Metody wykorzystane w tym badaniu to Adaptive Synthetic Multinomial Naive Bayes, Adaptive Synthetic k-nearest neighbours i Adaptive Synthetic Random Forest. Metoda Adaptive Synthetic służy do obsługi niezrównoważonych danych. Dane wykorzystane w tym badaniu to argumenty publiczne według prowincji w Indonezji. Wyniki uzyskane w tym badaniu to negatywne nastroje, które dominują we wszystkich prowincjach Indonezji. Istnieje związek między negatywnymi nastrojami a poziomem wykształcenia, korzystaniem z Internetu i wskaźnikiem rozwoju społecznego. Adaptive Synthetic Multinomial Naive Bayes działała lepiej niż inne metody, z dokładnością 0,882. Najwyższa dokładność metody Adaptive Synthetic Multinomial Naive Bayes wynosi 0,990 w prowincji Papua Barat.*

Słowa kluczowe: adaptacyjna synteza, klasyfikacja, dane dotyczące nierównowagi, dokładność

Introduction

Humans are created to interact with each other. Besides that, another essential human trait is to respond to a phenomenon around them [8]. Phenomena that are often answered to by society are phenomena that have a direct impact on their lives, for example, primary needs. The community often discusses economic, political, environmental, and humanitarian issues. Moreover, social media is growing very rapidly, so there are more and more platforms for people to express their opinions. Twitter is one of the favourite social media for people to express their views [2].

The expression of public opinion is closely related to the level of education, internet access, and human development index in a country, as well as Indonesia. People's critical thinking skills are trained in education and commenting on a phenomenon [37]. Moreover, education is strongly correlated with the human development index because education is one of the variables in the human development index. Internet access is crucial in the digital age, as most information is online.

The increase in fuel oil (BBM) became a topic of discussion among Indonesians in September 2022 [29]. Fuel is a primary need for most Indonesians. The mobility of Indonesian people is very high, and Fuel is crucial for them. Public comments about the fuel increase were conveyed on social media, and it became a trending topic. Twitter is the social media chosen by the public to express their opinions.

Public sentiment can be classified as positive or neutral, or negative. Sentiment analysis on fuel price increases needs to be done to determine whether people's opinions are classified as positive, neutral, or negative. In addition, the sentiment analysis results can be used as input for the government in taking and evaluating a policy.

In recent decades, Twitter sentiment analysis has used machine learning techniques to classify sentiment into positive, neutral, and negative classes [23]. Sentiment classification is crucial for the government in deciding future policies. A comparison of classification methods between logistic regression, Naive Bayes, Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD) were carried out to obtain the best way to classify Bengali book reviews, Multinomial Naive Bayes (MNB) being the best method with 84% accuracy [16]. The MNB method was applied by Rahman et al. [32] to classify sentiment in Bengali films, and the technique gave an accuracy of 86%. Several other studies on sentiment classification using the MNB method [7, 10, 32, 35, 38].

K-Nearest Neighbors (k-NN) is a machine learning method that uses distance techniques in classifying. Zamzuri et al. [41] use the k-NN method to classify the emotions contained in the text. The results obtained are the k-NN method which is capable of classifying with an accuracy of 79%. Hotel reviews in e-commerce applications using k-NN produce an accuracy of 87% [11].

Besides the MNB and k-NN methods, the Random Forest (RF) method accurately classifies text. The RF method produces better accuracy than the Support Vector Machine (SVM) and logistic regression in classifying feedback and reviewing airline services [31]. An accuracy of 96.42% was obtained by the RF method in classifying public opinion about sexual harassment cases from Brazilian anesthesiologists [5].

Conventional machine learning methods have the limitation of classifying datasets with unbalanced class distribution [40]. In addition, unbalanced data can reduce the accuracy of the classification method because there will be a prediction bias toward the majority class [19]. The phenomenon of fuel price increases in Indonesia has the potential to experience unbalanced data. The frequency of negative sentiment on the fuel increase will be more than neutral and positive.

Furthermore, methods of handling imbalanced datasets are needed; one such method is the Adaptive Synthetic (ADASYN) sampling approach [40]. The ADASYN algorithm works by generating synthetic data from minority classes. In the research of Khan et al. [36], the ADASYN method was able to increase the accuracy of the Extreme gradient boosting multi-classifier from 93% (without ADASYN) to 95% (with ADASYN). This method shows that the classification method will increase accuracy if the data is balanced. Some research on the ADASYN method [6, 20, 42].

Furthermore, the main contribution of this research is to describe the sentiment of fuel price increases per province in Indonesia and opinions that will describe the level of education, internet facilities, and the Human Development Index. In addition, classification using the ADASYN Multinomial Naive Bayes (MNB), ADASYN k-Nearest Neighbors (k-NN), and ADASYN Random Forest (RF) methods will be applied to classify positive, neutral, and negative sentiments in each province in Indonesia. The best method will be evaluated with the accuracy value.

1. Material and methods

This research used multinomial naive bayes for classify sentiments of people in Indonesia about increasing fuel, the sentiments divided into three categories. The three categories are positive, negative, and neutral statements. The positive statement is agreement arguments with increasing fuel, while disagreement is represented by negative statements, and neutral is impartial statement.

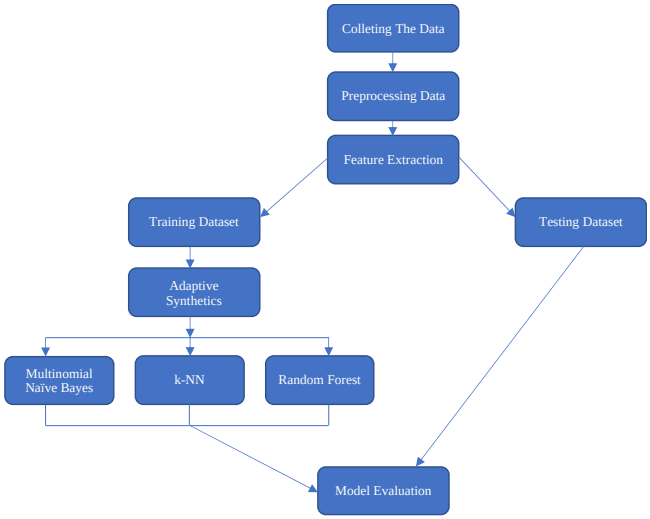

Fig. 1. The Research Step

The first step (figure 1) is collecting the data form twitter with the keywords "#BBMNaik", "BBM Naik", and "BBM", this step repeated for 34 provinces in Indonesia. Data collection in this study using the scraping method. Web scraping is a technique for automatically extracting information from various web documents. Get relevant content based on a query and convert unstructured format to structured representation [26]. The number of data retrieved in each province was as many as 100 tweets with the total tweets used in this study as many as 3400, where the tweets taken in Bahasa. The Sample of collected data is as follows in table 1.

Table 1. Sample of the web scraped data

Username	Tweet	User Location
@SilviaPutrii9	Penyesuaian bbm bertujuan untuk mengurangi beban subsidi besar akibat harga minyak dunia yang terus naik, kebijakan pemerintah sudah benar#BantuanBBMUntukRakyat #BLTBBMTepatSasaran #HematCermatBBM	Pontianak, Kalimantan Barat
@inthesky014	BBM naik itu biar yang sering bawa motor ke masjid pada jalan kaki, supaya pahalanya lebih banyak, subhanallah pemerintah kita ini sangat memperdulikan iman rakyatnya	Cianjur, Indonesia
@Kentringmanikk	@M45Broo_ Saya bisnis laundry dukung 1000% BBM naik. Bila perlu naikan terus, hilangkan subsidi BBM. Larang mobil pake BB fosil secepatnya. Subsidi bisa di pakai utk hal lain. Sudah tdk efisien negeri ini	Tanjung Emas, Indonesia

1.1. Text preprocessing

Results of collecting data was unstructured, so the next step is preprocessing data with Natural Language Processing (NLP) machine learning model [21]. Natural Language Processing have seven procedures, namely case folding, word normalization, cleansing filtering, stemming, and tokenizing.

a) Case Folding

The data is converted into the lowercase so that the uppercase and lowercase words with same meaning are not treated differently [27].

Table 2. Case Folding

Raw Data	Case Folding
Kenaikan harga BBM adalah hal sangat wajar, mengingat harga minyak dunia yang sedang melambung pesat #BantuanBBMUntukRakyat	kenaikan harga bbm adalah hal sangat wajar mengingat harga minyak dunia yang sedang melambung pesat #bantuanbbmuntukrakyat

b) Word Normalization

Word normalization is used to change words that are not standard as an informal word or shortened to a standard word in Bahasa.

Table 3. Word Normalization

Case Folding	Word Normalization
kenaikan harga bbm adalah hal sangat wajar mengingat harga minyak dunia yang sedang melambung pesat #bantuanbbmuntukrakyat	kenaikan harga bbm adalah hal sangat wajar mengingat harga minyak dunia yang sedang melambung pesat #bantuanbbmuntukrakyat

c) Cleansing

Cleansing is used to clean words that are not required such as hashtag (#), website address, username (@username), numbers, emojis and emails.

Table 4. Cleansing

Word Normalization	Cleansing
kenaikan harga bbm adalah hal kenaikan harga bbm adalah hal sangat wajar mengingat harga sangat wajar mengingat harga minyak dunia yang sedang minyak dunia yang sedang melambung pesat melambung pesat #bantuanbbmuntukrakyat	kenaikan harga bbm adalah hal sangat wajar mengingat harga minyak dunia yang sedang melambung pesat

d) Filtering

There are some words in the tokenized text that do not relate to any important concept or result, but may have important implications for the classifier. It is better to delete such words in advance.

Table 5. Filtering/Stopwords Removal

Cleansing	Filtering/Stopword Removal
kenaikan harga bbm adalah hal sangat wajar mengingat harga minyak dunia yang sedang melambung pesat	kenaikan harga wajar harga minyak dunia melambung pesat

e) Stemming

This step is to find the roots of words with deletion suffixes.

Table 6. Stemming

Filtering/Stopword Removal	Filtering/Stopword Removal
kenaikan harga wajar harga minyak dunia melambung pesat	naik harga wajar harga minyak dunia lambung pesat

f) Tokenization

Tokenization refers to the process of converting any text into a series of tokens, each distinct and independent of the other.

Table 7. Tokenization

Token
naik
harga
wajar
harga
minyak
dunia
lambung
pesat

1.2. Feature extraction

The process of converting text data to numbers is called Feature Extraction from text. This is also called text vectorization. The text contained in these tweets is unstructured, so in order to process it, it first needs to be pre-processed, six pre-processing techniques are used, and then features are extracted from the pre-processed data [1]. One of the feature extraction methods is TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is an algorithmic method useful for calculating the weight of any commonly used word This method is also known to be efficient, simple and gives accurate results. This approach would calculate the TF and IDF values of each token (word) in each document of the corpus. Generally, the TF-IDF method is used to find out how many times a word occurs in a document [34]. The calculation to find the TF-IDF value is as follows:

$$TF_{t,d}IDF_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

where $TF_{t,d}IDF_{t,d}$ is the weight of the term (t_j) to the documents (d_i). The $tf_{t,d}$ value is the frequency term (t) in document (d). However, if the term is not included in the document, the weight is zero [18]. Calculate the idf_t value:

$$idf_t = log_e \frac{1+n}{1+df_t} + 1 \quad (2)$$

where n is the number of whole documents in the collections while $\mathrm{df_t}$ is that of documents containing term (t).

1.3. Multinomial Naive Bayes (MNB)

Multinomial Naive Bayes (MNB) models the distribution word in the document as a multinomial. A document is treated as a sequence of words and assumed that each word position is generated independently of each other [37]. Multinomial Naive Bayes Classifier can be formulated as follows:

A tweet 'n' being of polarity 'p' is calculated as [30]:

$$P(p|n) \propto P(p) \prod_{1<k<nd} P(t_k|p) \quad (3)$$

where $P(t_k|p)$: represents the conditional probability that whether the term t_k occurs in a tweet of polarity p which is calculated as follows:

$$P(t_k|p) = \frac{count(t_k|p)+1}{count(t_p)+|V|} \quad (4)$$

Here, count ($t_k|p$) means the number of times the term t_k occurs in the tweets which have polarity p and count (t_p) means the total number of tokens present in the tweets of polarity p. Also, 1 and $|V|$ are added as smoothing constants which are added to avoid the mishaps in the calculation when the term does not occur at all in the tweets or the tweets is empty or null. This concept is better known as Laplace Smoothing [3]. $|V|$ is the number of terms in the total vocabulary of tweets.

$P(p)$: represents the prior probability of tweets being of polarity p which is calculated as follows:

$$P(p) = \frac{Number\ of\ tweets\ of\ polarity\ p}{Total\ number\ of\ tweets} \quad (5)$$

1.4. K-Nearest Neighbours (k-NN)

The k-NN classifier is a direct non-parametric classification algorithm [28]. A non-parametric method means that it does not make any assumptions about the distribution of the underlying data. Non-parametric algorithms such as k-NN use a flexible number of parameters, which often increases as more data becomes available [13].

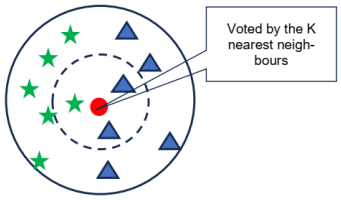

Fig. 2. Illustration of k-NN Algorithm

The k-NN algorithm is a classification algorithm that uses some K nearest data (neighbours) to determine a new data class. This algorithm classifies data based on their similarity or proximity to other data. To calculate the distance between two data in the k-NN algorithm using the Euclidean Distance method. The line length between points p and q is the Euclidean distance between them. If pi and qi are two locations in Euclidean n-space in Cartesian coordinates, then the distance from p to q is given by

$$d_E = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (6)$$

In general, the k-NN logarithm consists of three parts [4, 13]:

1. Determine the number of neighbours (K) used for class determination considerations.
2. Calculate the distance from the new data to each data point in the dataset.
3. Take several K data with the shortest distance, then determine the class of the new data.

1.5. Random Forest (RF)

Random Forest is a machine learning algorithm combining multiple decision trees' output to arrive at a single result [25]. Each tree in the Random Forest will issue class predictions. The class prediction with the most votes becomes the prediction candidate for the model. The greater the number of trees, the higher accuracy will result and prevent overfitting problems.

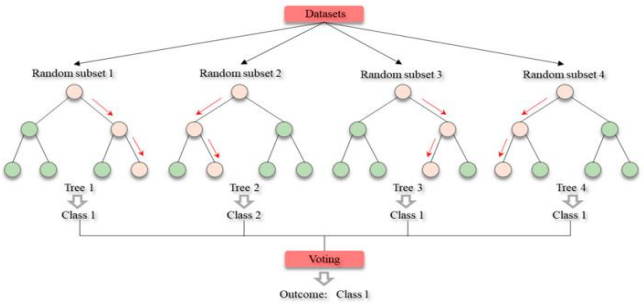

Fig. 2. Illustration of Random Forest Algorithm[39]

The Random Forest algorithm works can be described in the following steps [12].

1. Make a decision tree for each selected sample. Then the prediction results will be obtained from each decision tree that has been made.
2. A voting process is carried out for each prediction result. For classification problems, use the mode (the value that appears most often); for regression problems, use the mean (average value).
3. The algorithm will choose the prediction result with the most votes (most votes) as the final prediction.

1.6. Feature selection

Feature selection is the stage that helps reduce data size and remove features that don't important and improves accuracy [15]. Feature selection serves to reduce the size of the data dimensions as well as aims to select the best features from a feature data set. Feature selected by selecting important and relevant features to the data and reducing irrelevant features. The feature selection method used in this study is Chi Square. The chi square formula is as follows:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad (7)$$

where O_i is the observation value for category i and E_i is expected value for category i.

2. Results and discussion

We begin our analysis by looking at the spread of sentiment in Indonesia. The classification of sentiment in this study is divided into three categories, namely positive sentiment, neutral sentiment, and negative sentiment. The distribution of these sentiments illustrates the perspective of the Indonesian people towards rising BBM prices. In this case, we look at sentiment per province. Distribution can help local governments to make public policies.

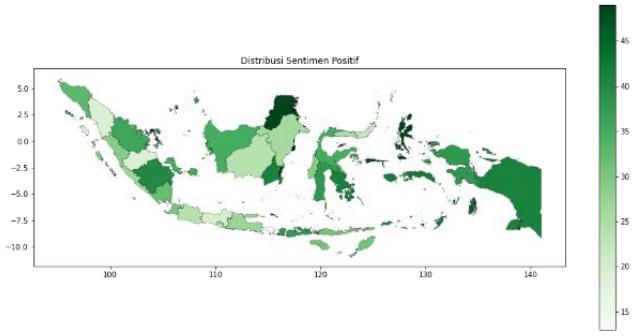

Fig. 4. Distribution of Positive Sentiment

Positive statements are argument that support the increasing BBM price. Based on figure 4, many people in the province disagree with the increase in fuel prices. This is indicated by a light map color. Meanwhile, the provinces of Kalimantan Utara and Maluku Utara were the provinces most supportive of the increase in fuel prices compared to other provinces.

The neutral statement illustrates that public opinion does not care about the increase in fuel prices. The provinces that distributed the most normal reports were Aceh, Jawa Tengah, and the D.I Yogyakarta (figure 5). The highest number of statements is 33. However, most provinces in Indonesia have neutral statements that are low on affiliation.

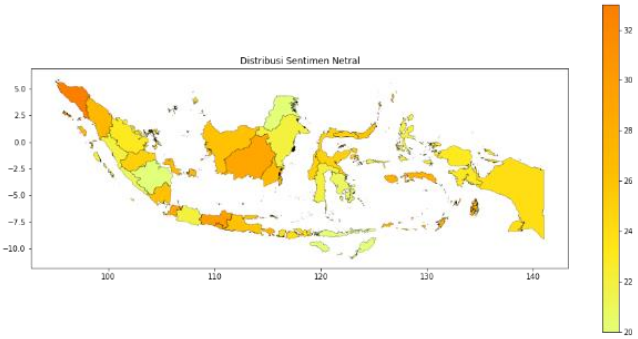

Fig. 5. Distribution of Neutral Sentiment

The majority of the general public reject the increase in fuel prices in Indonesia. This is indicated by the distribution of red in many parts of Indonesia. The province of Bali is the province with the highest negative statement with 58 sentiments. Other regions with a fairly strong red color are DKI Jakarta with 57 negative sentiments, Jambi with 56 negative sentiments and Jawa Barat with 54 negative sentiments. People in these areas often give their opinions in the form of news and suggestions so that these opinions are categorized as negative sentiments (figure 6).

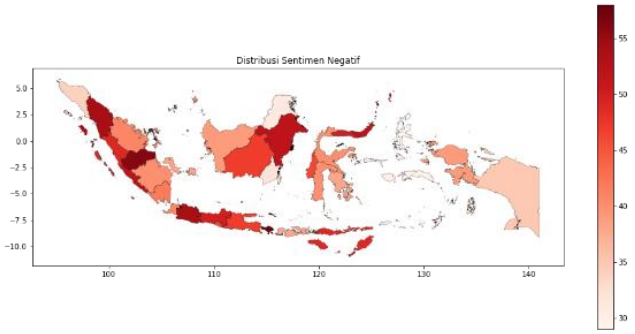

Fig. 6. Distribution of Negative Sentiment

The rejection was carried out because the fuel increase would hit people's purchasing power when salaries did not increase, so the additional allocation to buy fuel would increase and curtail other budgets so that people's purchasing power would decrease. In addition, according to news published by VOA Indonesia related to a national survey of Indonesian political indicators, most people reject the increase in subsidized fuel prices. The wide price gap between subsidized and non-subsidized fuel is considered ineffective in limiting subsidized fuel consumption. The survey results reveal that in terms of income, rejection of the fuel increase comes from those who earn more than Rp 4 million (54.2 percent) and people who make less than Rp1 million per month (69.1 percent). In terms of the type of work, the rejection of fuel increases generally comes from civil servants / private sector (82 percent) and groups of farmers, ranchers, and fishermen (67.3 percent).

Education shapes humans to think critically. Critical thinking is closely related to people expressing opinions on social media. People who express opinions in the media are educated people [9]. Several provinces with high levels of education wrote negative statement comments regarding the increase in fuel prices. It can be seen in figure 7 that DKI Jakarta has the highest level of education compared to other provinces in Indonesia. It is directly proportional to the number of negative comments about the fuel price increase.

The distributions above (figures 4–6) show positive, neutral, and negative reviews and show that the distribution of education levels is uneven. The level of education is directly proportional to the infrastructure (search the literature). The more infrastructure supporting education, the better the level of education in the area.

The word cloud results visualize the words that Twitter users coonly use to express their opinions. The bigger the font size on word cloud then the topic frequently discussed by Twitter users. The following figure shows the word cloud's results.

Based on figure 8 it can be seen positive sentiment, neutral sentiment, and negative sentiment along with the words used. This is useful to know the description of public sentiment in every province in Indonesia regarding the increase in fuel prices set by the government. The word "Harga" dominates in the neutral and negative sentiments. The difference lies in the next term. The negative sentiment contains a sentence that protests. The distribution of positive, neutral and negative sentiment in other provinces has the same pattern.

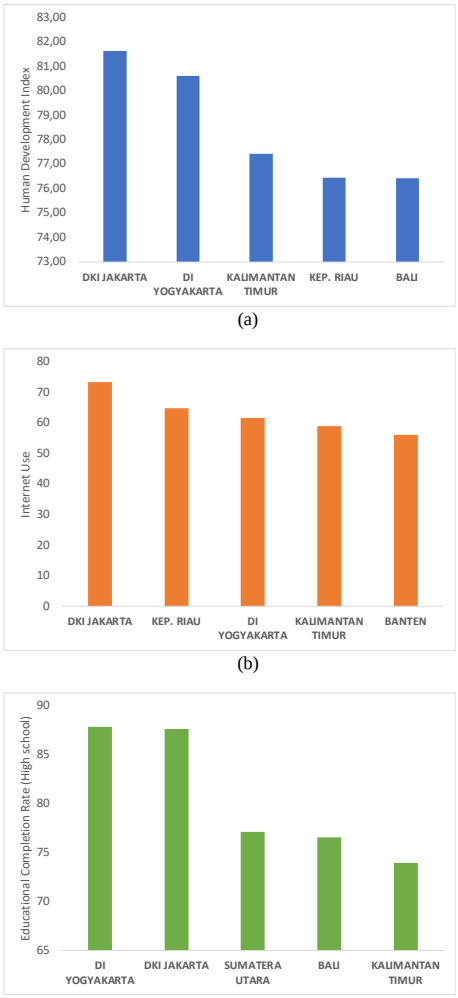

Fig. 7. Variables that describe the level of education (the highest 5 provinces): (a) Human Development Index, (b) Internet Use, (c) Educational Completion Rate (Heigh School)

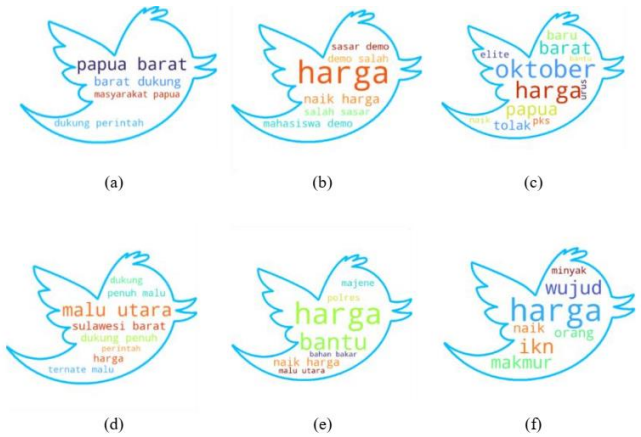

Fig. 8. Sample Wordcloud Sentiment Increase in fuel prices: (a) Papua Barat Positive, (b) Papua Barat Neutral, (c) Papua Barat Negative, (d) Sulawesi Barat Positive, (e) Sulawesi Barat Neutral, (f) Sulawesi Barat Negative

Public response to a particular phenomenon cannot be uniform. If the phenomenon is positive, the frequency of positive comments on social media is more than negative comments, as well as on negative phenomena. See figure 9 (a), the amount of data in each category is disproportionate. The amount of data in the negative category tends to be higher when compared to the neutral category and the positive category. This is imbalance data class.

Imbalanced class data is a common problem in classification machine learning where there is a disproportionate ratio of each class. Most machine learning algorithms don't work very well with imbalanced class data. If this is allowed, then the machine learning model that is built will tend to predict data into the majority data class.

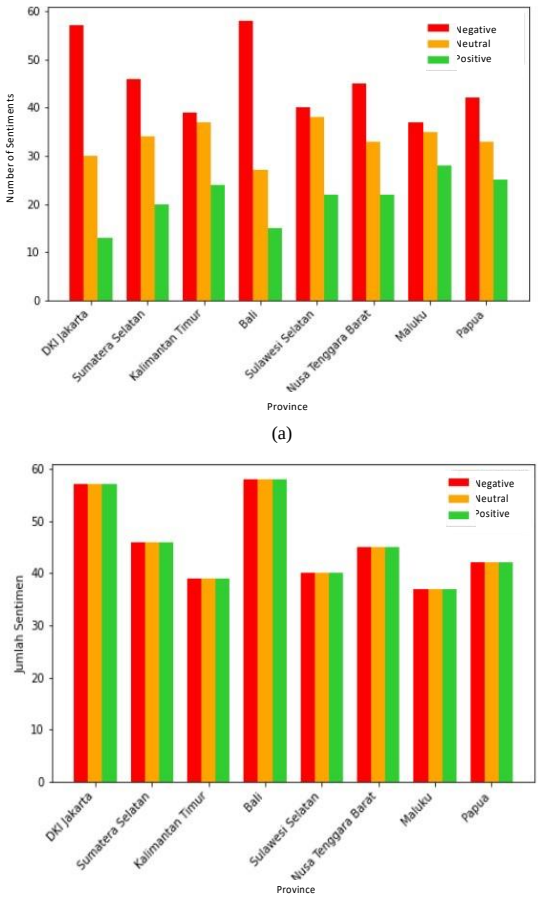

Fig. 9. Before and After Applied ADASYN: (a) before, (b) after

The Adaptive Synthetic (ADASYN) method is used to overcome this problem [14, 17, 22, 24]. After applying ADASYN (figure 9(b)), the data in each category becomes proportional. The data distribution in each category is even so that the data is ready to be used to build machine learning models. Machine learning models are built using Multinomial Naive Bayes (MNB), k-Nearest Neighbours (k-NN), and Random Forest (RF) algorithms.

The classifier model that has been built is used to test data testing. Each data is predicted using the MNB, k-NN, and RF algorithms to be classified into three sentiment categories: positive, neutral, or negative. In this study, researchers tested 680 tweets (data testing) related to rising fuel prices.

Evaluation of results using the confusion matrix to determine the level of accuracy. Based on the data testing process, obtained the level of accuracy in each province in Indonesia is shown in the table below.

Based on table 8, the accuracy of the MNB algorithm is better than the k-NN and RF methods, with an average accuracy of 0.882 (the highest compared to other methods). In general, the MNB accuracy in each province is the highest compared to the k-NN and RF methods. The highest accuracy is in Papua Barat Province, with an accuracy value of 0.990, while the lowest is in Sulawesi Utara Province. The high accuracy value is because the testing data has the vocabulary contained in the training data, so the model can learn it based on the training data and adequately process the testing data.

Table 8. Accuracy

Province	Accuracy		
	MNB	K-NN	RF
Aceh	0.905	0.95	0.80
Bali	0.956	0.90	0.95
Bangka Belitung	0.970	0.80	0.95
Banten	0.960	0.95	0.80
Bengkulu	0.806	0.80	0.75
DI. Yogyakarta	**0.980**	0.95	0.85
DKI Jakarta	0.882	0.90	0.95
Gorontalo	0.942	0.85	0.80
Jawa Barat	0.930	0.95	0.85
Jambi	0.970	0.90	0.80
Jawa Tengah	0.980	0.90	0.90
Jawa Timur	0.966	0.95	0.90
Kalimantan Barat	0.921	0.90	0.90
Kalimantan Selatan	0.980	0.95	0.90
Kalimantan Tengah	0.941	0.85	0.95
Kalimantan Timur	0.900	0.75	0.80
Kalimantan Utara	0.667	0.85	0.95
Kep. Riau	0.933	0.80	0.90
Lampung	0.857	0.80	0.85
Maluku	0.818	0.80	0.70
Maluku Utara	0.957	0.75	0.80
Nusa Tenggara Barat	0.960	0.90	0.90
Nusa Tenggara Timur	0.817	0.80	0.80
Papua	0.814	0.85	0.95
Papua Barat	0.990	0.95	0.95
Riau	0.671	0.85	0.75
Sulawesi Barat	0.986	0.95	0.80
Sulawesi Selatan	0.775	0.90	0.85
Sulawesi Tengah	0.833	0.90	0.85
Sulawesi Tenggara	0.897	0.90	0.80
Sulawesi Utara	0.571	0.65	0.70
Sumatera Barat	0.967	0.90	0.80
Sumatera Selatan	0.750	0.75	0.80
Sumatera Utara	0.750	0.85	0.75
Average Accuracy	**0.882**	0.860	0.850

3. Conclusion

Based on the results of this study it can be seen that the province with a dense population such as DKI Jakarta, Jawa Barat, and Bali tend to resist the increase in fuel prices. The majority of people on each province uses the word “harga” in giving its opinion regarding the increase in fuel prices either negative sentiment, neutral sentiment, or positive sentiment. There is a relationship between negative sentiment and the human development index, internet use, and educational completion rate (high school). The performance of the machine learning model that has been built is measured using accuracy. The ADASYN method is applied to balance the minority and majority classes. The ADASYN Multinomial Naive Bayes (MNB) method has better performance than the ADASYN k-Nearest Neighbors (k-NN) and ADASYN Random Forest (RF) methods. The average accuracy obtained by the MNB method is 0.882 or 88.2%. Near-perfect accuracy values are in the province of Papua Barat. You can use the hybrid method and add testing data to improve performance in further research.

References

[1] Ahuja R. et al.: The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science 152, 2019, 341–348 [http://doi.org/10.1016/j.procs.2019.05.008].

[2] Ali H. et al.: Deep Learning-Based Election Results Prediction Using Twitter Activity. Soft Computing 26(16), 2022, 7535–43 [http://doi.org/10.1007/s00500-021-06569-5].

[3] Amity U. et al.: Abstract Proceedings of International Conference on Automation, Computational and Technology Management (ICACTM-2019), 2019.

[4] Andrian R. et al.: K-Nearest Neighbor (k-NN) Classification for Recognition of the Batik Lampung Motifs. Journal of Physics: Conference Series 1338(1), 2019 [http://doi.org/10.1088/1742-6596/1338/1/012061].

[5] Asian J. et al.: Sentiment Analysis for the Brazilian Anesthesiologist Using Multi-Layer Perceptron Classifier and Random Forest Methods. Journal Online Informatika 7(1), 2022, 132 [http://doi.org/10.15575/join.v7i1.900].

[6] Balaram A., Vasundra S.: Prediction of Software Fault-Prone Classes Using Ensemble Random Forest with Adaptive Synthetic Sampling Algorithm. Automated Software Engineering 29(1), 2021, 6 [http://doi.org/10.1007/s10515-021-00311-z].

[7] Budiawan Zulfikar W. et al.: Sentiment Analysis on Social Media Against Public Policy Using Multinomial Naive Bayes. Scientific Journal of Informatics 10(1), 2023 [http://doi.org/10.15294/sji.v10i1.39952].

[8] Bustillos A. et al.: Approaching Dehumanizing Interactions: Joint Consideration of Other-, Meta-, and Self-Dehumanization. Current Opinion in Behavioral Sciences 49, 2023, 101233 [http://doi.org/10.1016/j.cobeha.2022.101233].

[9] Eberwein T.: ‘Trolls’ or ‘Warriors of Faith’?: Differentiating Dysfunctional Forms of Media Criticism in Online Comments. Journal of Information, Communication and Ethics in Society 18(1), 2020, 131–143 [http://doi.org/10.1108/JICES-08-2019-0090].

[10] Farisi A. A. et al.: Sentiment Analysis on Hotel Reviews Using Multinomial Naive Bayes Classifier. Journal of Physics: Conference Series 1192(1), 2019 [http://doi.org/10.1088/1742-6596/1192/1/012024].

[11] Gazali Mahmud F. et al.: Implementation Of K-Nearest Neighbor Algorithm With SMOTE For Hotel Reviews Sentiment Analysis. Sinkron: Jurnal Dan Penelitian Teknik Informatika 8(2), 2023, 595–602 [http://doi.org/10.33395/sinkron.v8i2.12214].

[12] Ghosh D., Cabrera J.: Enriched Random Forest for High Dimensional Genomic Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 19(5), 2022, 2817–2828 [http://doi.org/10.1109/TCBB.2021.3089417].

[13] Hasdyna N. et al.: Improving the Performance of K-Nearest Neighbor Algorithm by Reducing the Attributes of Dataset Using Gain Ratio. Journal of Physics: Conference Series 1566(1), 2020 [http://doi.org/10.1088/1742-6596/1566/1/012090].

[14] He H. et al.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, 1322–1328 [http://doi.org/10.1109/IJCNN.2008.4633969].

[15] Herhianto A.: Sentiment Analysis Menggunakan Naive Bayes Classifier (Nbc) Pada Tweet Tentang Zakat. 2020.

[16] Hossain E. et al.: Sentiment Polarity Detection on Bengali Book Reviews Using Multinomial Naive Bayes. Progress in Advanced Computing and Intelligent Engineering (ed.Chhabi Rani Panigrahi et al.), Springer Singapore, 2021, 281–292.

[17] Hu Z. et al.: A Novel Wireless Network Intrusion Detection Method Based on Adaptive Synthetic Sampling and an Improved Convolutional Neural Network. IEEE Access 8, 2020, 195741–195751 [http://doi.org/10.1109/ACCESS.2020.3034015].

[18] Jalilifard A. et al.: Semantic Sensitive TF-IDF to Determine Word Relevance in Documents, 2020 [http://doi.org/10.1007/978-981-33-6977-1].

[19] Jiang C. et al.: Benchmarking State-of-the-Art Imbalanced Data Learning Approaches for Credit Scoring. Expert Systems with Applications 213, 2023, 118878 [http://doi.org/10.1016/j.eswa.2022.118878].

[20] Koh J. E. W. et al: Automated Classification of Attention Deficit Hyperactivity Disorder and Conduct Disorder Using Entropy Features with ECG Signals. Computers in Biology and Medicine 140, 2022, 105120 [http://doi.org/10.1016/j.compbiomed.2021.105120].

[21] Kurniasih A., Lindung P. M.: On the Role of Text Preprocessing in BERT Embedding-Based DNNs for Classifying Informal Texts. International Journal of Advanced Computer Science and Applications 13(6), 2022, 927–934 [http://doi.org/10.14569/IJACSA.2022.01306109].

[22] Kurniawati Y. E. et al.: Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data. 2018 4th International Conference on Science and Technology (ICST), 2018, 1–6 [http://doi.org/10.1109/ICSTC.2018.8528679].

[23] Leelawat N. et al.: Twitter Data Sentiment Analysis of Tourism in Thailand during the COVID-19 Pandemic Using Machine Learning. Heliyon 8(10), 2022, e10894 [http://doi.org/10.1016/j.heliyon.2022.e10894].

[24] Liu J. et al.: A Fast Network Intrusion Detection System Using Adaptive Synthetic Oversampling and LightGBM. Computers & Security 106, 2021, 102289 [http://doi.org/10.1016/j.cose.2021.102289].

[25] Liu Y., Wu H.: Prediction of Road Traffic Congestion Based on Random Forest. 2017 10th International Symposium on Computational Intelligence and Design (ISCID) 2, 2017, 361–364 [http://doi.org/10.1109/ISCID.2017.216].

[26] Lytvyn V. et al.: Identifying Textual Content Based on Thematic Analysis of Similar Texts in Big Data. 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT) 2, 2019, 84–91 [http://doi.org/10.1109/STC-CSIT.2019.8929808].

[27] Mayo M.: A General Approach to Preprocessing Text Data, 2017.
[28] Moosavian A. et al.: Comparison of Two Classifiers; K-Nearest Neighbor and Artificial Neural Network, for Fault Diagnosis on a Main Engine Journal-Bearing. Shock and Vibration 20(2), 2013, 263–272 [http://doi.org/10.3233/SAV-2012-00742].
[29] Nadhifah D. et al.: Analysis of the Impact of the Increase in Fuel Oil (BBM) on Household Economic Activities. Journal of Contemporary Gender and Child Studies (JCGCS) 1(1), 2022 [https://zia-research.com/index.php/jcgcs].
[30] Nazrul Syed S.: Multinomial Naive Bayes Classifier for Text Analysis (Python). Towards Data Science, 2018.
[31] Patel A. et al.: Sentiment Analysis of Customer Feedback and Reviews for Airline Services Using Language Representation Model. Procedia Computer Science 218, 2023, 2459–2467 [http://doi.org/10.1016/j.procs.2023.01.221].
[32] Rahman R. et al.: Sentiment Analysis on Bengali Movie Reviews Using Multinomial Naive Bayes. 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021, 1–6 [http://doi.org/10.1109/ICCIT54785.2021.9689787].
[33] Rennie J. D. M. et al.: Tackling the Poor Assumptions of Naive Bayes Text Classifiers, 2003.
[34] Ridho Lubis A. et al.: The Effect of the TF-IDF Algorithm in Times Series in Forecasting Word on Social Media. Indonesian Journal of Electrical Engineering and Computer Science 22(2), 2021, 976 [http://doi.org/10.11591/ijeecs.v22.i2.pp976-984].
[35] Sahib N. G. et al.: Sentiment Analysis of Social Media Comments in Mauritius. IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), 2023, 860–865 [http://doi.org/10.1109/CCWC57344.2023.10099291].
[36] Salauddin Khan M. et al.: Comparison of Multiclass Classification Techniques Using Dry Bean Dataset. International Journal of Cognitive Computing in Engineering 4, 2023, 6–20 [http://doi.org/10.1016/j.ijcce.2023.01.002].
[37] Solikah M., Dian N.: The Effectiveness of the Guided Inquiries Learning Model on the Critical Thinking Ability of Students. Jurnal Pijar Mipa 17(2), 2022, 184–191 [http://doi.org/10.29303/jpm.v17i2.3276].
[38] Surya P. P. et al.: Analysis of User Emotions and Opinion Using Multinomial Naive Bayes Classifier. 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, 410–415 [http://doi.org/10.1109/ICECA.2019.8822096].
[39] Yang J. et al.: Delineation of Urban Growth Boundaries Using a Patch-Based Cellular Automata Model under Multiple Spatial and Socio-Economic Scenarios. Sustainability (Switzerland) 11(21), 2019 [http://doi.org/10.3390/su11216159].
[40] Yu B. et al.: Classification Method for Failure Modes of RC Columns Based on Class-Imbalanced Datasets. Structures 48, 2023, 694–705 [http://doi.org/10.1016/j.istruc.2022.12.063].
[41] Zamsuri A. et al.: Classification of Multiple Emotions in Indonesian Text Using The K-Nearest Neighbor Method. Journal of Applied Engineering and Technological Science (JAETS) 4(2), 2023, 1012–1021 [http://doi.org/10.37385/jaets.v4i2.1964].
[42] Zhai J. et al.: Binary Imbalanced Data Classification Based on Diversity Oversampling by Generative Models. Information Sciences 585, 2022, 313–43 [http://doi.org/10.1016/j.ins.2021.11.058].

M.Sc. Fatkhurokhman Fauzi
e-mail: fatkhurokhmanf@unimus.ac.id

He is a lecturer in the statistics department. Research focus: data mining, text mining, machine learning, forecasting, environmental statistics, and climate modeling. He is currently a researcher in the Degrees Initiative research group.

http://orcid.org/0000-0002-8277-8638

B.Sc. Ismatullah
e-mail: ismatullahp17@gmail.com

Research focus: data mining, text mining, and machine learning.

http://orcid.org/0009-0005-7472-1761

M.Sc. Indah Manfaati Nur
e-mail: indahmnur@unimus.ac.id

She is a lecturer in the statistics department. Research focus: text mining and applied statistics.

http://orcid.org/0000-0002-1017-7323