

Social Media Text Analysis on Public's Sentiments of Covid-19 Booster Vaccines

Yohan Kristian[1], Adira Valdi Yesenia[1], Safina Safina[1], Anindya Apriliyanti Pravitasari[1], Eka Novita Sari[2], and Tutut Herawan[3,4(✉)]

[1] Departement of Statistics, Universitas Padjadjaran, Jl. Ir. Soekarno KM. 21, Jatinangor, Sumedang, West Java 45363, Indonesia
{yohan19001,adira19001,safina19001}@mail.unpad.ac.id, anindya.apriliyanti@unpad.ac.id

[2] AMCS Research Center, Sleman, Yogyakarta, Indonesia
eka@amcs.co

[3] Sekolah Tinggi Pariwisata Ambarrukmo Yogyakarta, Jl. Ringroad Timur No. 52, Bantul, Daerah Istimewa Yogyakarta 55198, Indonesia

[4] Institute for Big Data Analytics and Artificial Intelligence, UiTM Shah Alam, Selangor Darul Ehsan, 40450 Shah Alam, Malaysia
tutut@um.edu.my

Abstract. Two years of the COVID-19 Pandemic, countries across the world have started the process of vaccination in two-step doses. WHO stated that six months after the second dose injection, the effectiveness of the EUL (emergency use listing) vaccines has decreased by about 8%. Therefore, booster vaccines are recommended to be developed. Indonesia launched booster vaccinations with the objective of restoring decreased immunity and giving clinical protection. This study assesses the opinions of Indonesian citizens regarding the booster vaccine through social networks (Twitter and Youtube), which are mined through the Twitter API and Python Selenium Web Driver. Several algorithms have been employed to evaluate the best predictions of public sentiment. Each of them is given four scenarios to handle the imbalanced data: not handling the imbalance, and handling it with SMOTE, random oversampling and random undersampling. Support Vector Machines, Random Forest, Bidirectional Recurrent Neural Network, Gaussian Naive Bayes, Logistic Regression, Bernoulli Naive Bayes, and CatBoost Classifiers are executed under the same experimental setup. The best performance is given by CatBoost with ROS for handling the imbalance data; the accuracy is 88%, the weighted average f1-score is 88%, while the precision and recall averages are 89% and 88%, respectively.

Keywords: sentiment analysis · COVID-19 · booster vaccine · social media · text analysis

1 Introduction

The first case of Covid-19 was reported in late 2019 in Wuhan, China, and since then WHO has recorded over 500 million positive cases around the world [1]. The first action many countries took to control the virus' outbreak was to implement social distancing,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14104, pp. 209–224, 2023.
https://doi.org/10.1007/978-3-031-37105-9_15

national border closing and quarantine. It quickly becomes clear that these regulations are not enough to contain the virus that has killed more than 156,673 people in Indonesia alone [1]. There needs to be another strategy to protect citizens, particularly health-care workers, seniors and those with underlying health conditions [2]. To meet the imminent need for stronger protection, a new pandemic vaccine development paradigm has been proposed [3]. Although the need for vaccines is widely accepted all over the world, the development of Covid-19 vaccines is not without its challenges: such as deciding over the best approach to creating the vaccines, side effects that may occur, testing strategies, unknown duration of immunity, and last but not least, funds. Finally, a large demand for vaccines will arise due to the pandemic. Thus, clinical and serologic investigations will be needed to confirm which are still at risk once vaccinations are accessible, and studies that could serve as the foundation for developing a globally equitable vaccine-allocation system [3].

In Presidential Decree No. 18 of 2020, which was stipulated on September 3, 2020, President of Republic Indonesia, Joko Widodo, developed the National Team for Accelerating the Development of Coronavirus Disease 2019 (Covid-19). The team has 4 goals: first to accelerate the development of the Covid-19 vaccine in Indonesia, second to realize the national resilience and independence of the nation in the development of the Covid-19 vaccine, third to increase the synergy of research, development, study, and application of science and technology, as well as inventions and innovations, production, distribution, and use and/or utilization of the Covid-19 vaccine between the government and science and technology institutions as well as technological resources in Covid-19 vaccine development, and finally to utilized and increased capacity, as well as national capabilities in the development of the Covid-19 vaccine [4]. After a rigorous process of testing in clinical trials to prove that they meet the standards of safety and effectiveness, scientists were able to manufacture safe and efficacious Covid-19 vaccines in record speed [5]. The first vaccination in Indonesia took place on 31 January, 2021 with the President of Republic Indonesia Joko Widodo, as the first person who was vaccinated with a dose of Sinovac vaccine [6]. To date, 66,3% of the world has received at least one dose of a Covid-19 vaccine [7]. In Indonesia, over 201 million have received the first dose, and over 168 million have received the second dose of Covid-19 vaccines [8]. Nonetheless, COVID-19 immunization efforts must continue to focus on reducing death and serious diseases, as well as protecting the health-care system [9].

Despite the growing number of vaccinated citizens around the world, a few nations face new challenges, such as viral variant emergence and concerns about fading protection following immunization [10]. Mizrahi et al., looked at data from Israel's Ministry of Health and Maccabi Healthcare Services (MHS). They discovered that with time, protection against infection and disease deteriorated (from December 2020 to July 2021). People who were vaccinated in January and February, for example, were 53 percent more likely than those who were vaccinated in March and April, to test positive for COVID-19 [11]. WHO stated that six months after the second dose injection, the effectiveness of the EUL (emergency use listing) vaccines has decreased by about 8% over a period of 6 months in all age groups [9]. Those, and several other research determined the need for a third vaccine dose. Indonesia has started their third vaccine dose program on Wednesday, 12 January, 2022. For the third dose of vaccines, it can be done with the same type

of vaccine (homologous) or by giving a different vaccine from the first and second dose (heterologous). The ministry of Health prioritizes health workers on the grounds of an emergency due to the Delta variant [12] invasion. And its current expansion has taken into account vaccine safety, efficacy and quality [13].

The news of the third dose of vaccination of Covid-19 in Indonesia has gotten a variety of responses from the citizens, particularly on social media. Facebook, Twitter and Instagram, for example, offer a lot of untapped potential for providing non-trivial knowledge and might be utilized to analyze changes in the social, public health, and economic arenas [14]. When a rapid decision is necessary, obtaining information directly from the public via social media is very beneficial, as traditional communication methods such as surveys and interviews do not allow this. This research aims to analyze the public's sentiment towards COVID-19 booster vaccination in Indonesia while also reviewing the most appropriate method of social media text analysis for this type of data. From a political perspective, the massive demand for political information can be regarded as an important factor. Thus, sentiment analysis plays an important role in national security and public opinion analysis [15]. The result of this research not only could play a part in the next decision-making process regarding COVID-19 vaccination by giving insight to the government into how Indonesia's citizens responded towards the third dose vaccination of COVID-19, but also highlight the best method of social media text analysis.

This research will fill the gap of the previous research from the total of data collected, method of data cleaning and analysis, and the programming language used in analysis. In the research done by Khoiril Hikmah, the data collected from Twitter with a total of 100 tweets are used to analyze the sentiment using Naive Bayes Algorithm with PHP programming language [16]. The data used in this research is collected by Rapid Miner from Twitter tweets and Youtube comments, containing 647 neutral comments, 374 negative comments, and 282 positive comments. Each of them is given four scenarios to handle the imbalanced data, which are: not handling the imbalance, and handling it with random oversampling (ROS), random undersampling (RUS), and synthetic minority over-sampling technique (SMOTE). The method compared to analyze the sentiments on this research are CatBoost Classifier, Support Vector Machine (SVM), Random Forest, Logistic Regression, GaussianNB, BernoulliNB, and Bidirectional Gated Recurrent Unit (BiGRU) method using python programming language.

Further research in this topic can be done to fill the gap that this research can not reach. Since this research only analyzes the sentiment of Indonesia's citizens on social media, the result of this research only applies to the situation in Indonesia and can not in any way be generalized to other countries or territories. Moreover, the result of this research can be improved by a larger data set collected from other sources, a more thorough cleaning process of the data and the use of a more standard language on the data collection.

The remainder of the paper is organized as follows. In Sect. 2, more comparable studies are provided and discussed in order to lead to the research question. Section 3 details the methodology, including the tools used to gather tweets and comments, as well as to support the method employed in this research. For Sect. 4, the findings are

presented and discussed, and in Sect. 5, we conclude this research and highlight future research.

2 Related Works

Using unigram and bigram (*n*-gram) and supervised learning with simple Support Vector Machines (SVM), research of Balahur [17] was written, with the topic of emotional analysis of Twitter datasets. On the paper result, it is concluded that on one hand, the best method to be used for emotional analysis is a combined form of unigram and bigram. Also by using unique tags, such as emotive words and modifiers, are seen to strongly improve the performance rating of emotions. (joy, happiness, sadness, fear, and etc.) In another study by Narasamma and Sreedevi [18], they introduced a hybrid machine learning model named Catboost Recurrent Neural Framework (CRNF) with an error pruning mechanism to analyze the Twitter data based on user opinion. Initially, a twitter-based dataset on the coronavirus COVID-19 vaccine is obtained, which is then pre-processed and trained to the system. Furthermore, the suggested CRNF model divides sentiments into three categories: positive, negative, and neutral. Furthermore, the sentiment analysis procedure is carried out using Python, and the parameters are determined.

Employing the Naive Bayes Algorithm method, Pristiyono, *et al.* [19] analyses sentiment analysis of COVID-19.The data crawling process in the article was done manually by using the access token received from the Twitter API using the rapid miner tools to extract the requested information and data. Hikmah, *et al.* [16] wrote another paper about vaccine booster sentiment analysis. The dataset was retrieved from Twitter about Indonesian netizen perspective towards the booster vaccine. Case folding, tokenizing, filtering, and stemming are all part of the data preprocessing stage. While the sentiment classification is divided into positive sentiment, negative emotion and neutral sentiment by using the Naive Bayes algorithm. Another study about sentiment analysis on COVID-19 vaccine booster was written by Ong, Pauzi and Gan [14]. The data was gathered from Twitter users in Malaysia and identified three important latent Dirichlet allocation (LDA) topics: (1) type of vaccination booster; (2) effects of vaccination booster; (3) vaccination program operation.

This study builds on the related work by employing several known methodologies to determine the optimal model for predicting public sentiment towards the COVID-19 booster vaccine. The models compare each other with the imbalanced data that was previously handled. Table 1 summarizes the previous state of the art, which related to this research.

Table 1. Related work of sentiment analysis

References	Topic	Model	Accuracy (%)
Balahur [17]	Sentiment Analysis in Social Media Texts	SVM	79.96%
Narasamma and Sreedevi [18]	Twitter based Data Analysis in Natural Language Processing using a Novel Catboost Recurrent Neural Framework	CRNF	99.34%
Pristiyono, Ritonga, Ihsan, Anjar and Rambe [21]	Sentiment Analysis of COVID-19 Vaccine in Indonesia Using Naive Bayes Algorithm	Naive Bayes	-
Hikmah, Fauzan, and Harliana [11]	Sentiment Analysis of Vaccine Booster during Covid-19: Indonesian Netizen Perspective Based on Twitter Dataset	Naive Bayes	89.00%
Ong, Pauzi and Gan [13]	Text Mining and Determinants of Sentiments Towards the COVID-19 Vaccine Booster of Twitter Users in Malaysia	LDA	-

3 Material and Proposed Method

Several steps were done to achieve the results. After mining the data required to run the algorithms, the process of cleaning the data and labeling each comment was needed. A case of imbalance data was found after labeling each comment, therefore three methods to handle the imbalance were applied alongside the classification algorithms. The accuracy of each model combination was compared to evaluate which model is the best for the data. The specification and flow of analysis system is visualized in Fig. 1.

3.1 Data Processing

The proposed approach utilized the Twitter and YouTube comment data for analyzing the sentiments. Here, both of the data sources are based on COVID-19 booster vaccine, collected by using the keywords "vaccine" and "booster" in Indonesian language. The data crawling operation is done manually using the Twitter API access token and the Rapid miner tools to extract the desired information and data. YouTube comments from several videos reporting on booster vaccine were also mined using Rapid miner tools. The utilized dataset details about the 1282 numbers of tweets and YouTube comment altogether that involves the aftereffect of booster vaccine, information about the availability of the vaccine, and so on. In this, the collected data are based on the category of "positive" reviews, negative reviews and neutral reviews. Moreover, the collected dataset is given to the developed Catboost model for further processing. The comments that are categorized as negative comments are those containing bad words, ridicule, or

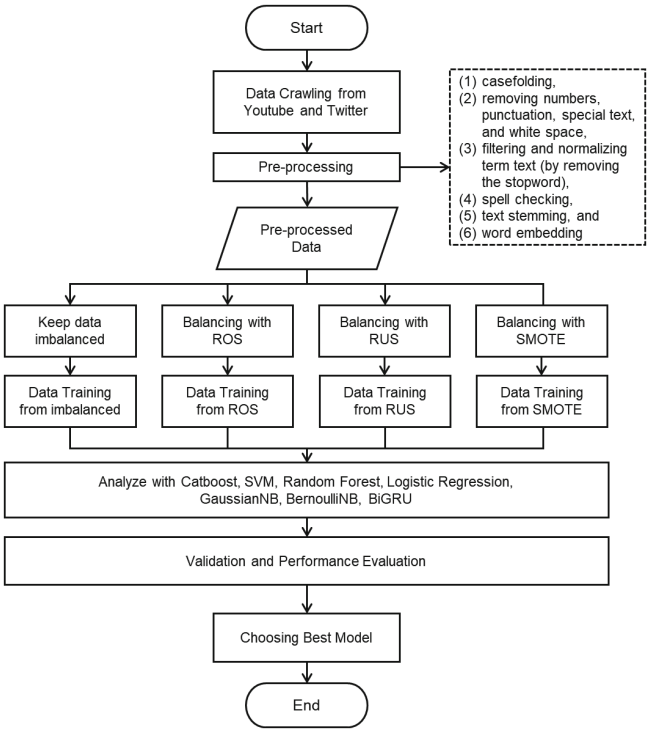


Fig. 1. The specification and flow of analysis system

opposition towards COVID-19 booster vaccine. The category of neutral comments are given if the comments only contain information or question without stating partiality. Finally, the comments that are deemed as positive comments contain praises, statements of agreement, even encouragement to others to participate in the third dose of COVID-19 vaccination. Table 2 below are examples of labeled data for this study.

Table 2. Examples of the categorized comments

Comments	Label/ Category
Saya vaksin 1,2, dan booster di Indonesia. Alhamdulillah taat aturan dan protokol kesehatan. Kita sehat, Indonesia kuat	Positive
Sementara itu, mekanisme Heterolog, yaitu pemberian vaksin booster dengan menggunakan jenis vaksin yang berbeda dengan vaksin primer dosis lengkap yang telah didapat sebelumnya	Neutral
Semua orang yang sudah booster dan taat prokes kena covid. Jadi apa gunanya vaksin dan prokesnya itu?	Negative

Figure 2 shows the number of comments each label has after the labeling process. It shows that the data is unbalanced, with the Positive category making up the minority and the Neutral category the majority.

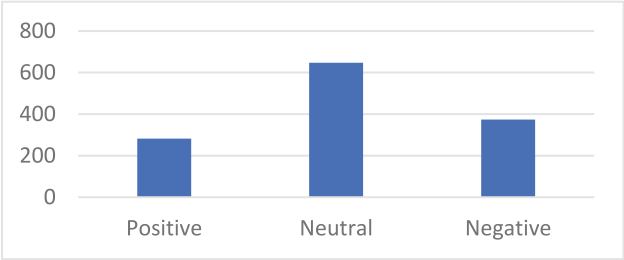


Fig. 2. Histogram of labeled comments

At the preprocessing stage, the data will be uniform in shape and format. This stage consists of (1) casefolding, (2) removing numbers, punctuation, special text, and white space, (3) filtering and normalizing term text (by removing the stopword), (4) spell checking, (5) text stemming, and (6) word embedding. This stage produces terms or basic words that are free from prefixes or suffixes. As part of preprocessing, the data that has been categorized were then labeled word by word. This labeling process is to identify the raw data that will provide context to it, and enable machine learning to use the data and learn from it. Table 3 are the results of the cleaned data that is used in this study.

Table 3. Examples of the cleaned comments

No	Label	Comment	Cleaned comment
1	0	['tinjau', 'kena', 'vaksi', boster', 'pfizer',…]	Tinjau kena vaksi boster Pfizer astrazeneca
2	0	['vaksin', 'booster', 'hanya', 'tahan', 'bulan',…]	Vaksin booster hanya tahan bulan telah apa
3	2	['saya', 'tunggu', 'efek', 'samping',…]	Saya tunggu efek samping vaksin booster kok
4	2	['saya', 'tidak', 'demam', 'sama', 'sekali',…]	Saya tidak demam sama sekali sampai suah sore

3.2 Methodology for Handling Imbalance Data

The condition of imbalanced data poses a risk to classification fairness because of uneven data representations in which the minority class is usually also an important one. Therefore, we require methods to improve its recognition rates by balancing the dataset and

reducing the risk of analysis or machine learning algorithms skewing toward the majority. Several methods for handling imbalance data are employ in this study, i.e. ROS, RUS, and SMOTE, that compared each other and also with data that remains imbalanced. The method of Random Over Sampling (**ROS**) is used to address data imbalances. Without increasing the variation of the class data, ROS replicates the minority class data sample and adds it to the training data set. The size of the training data set is increased by ROS by repeating the original sample in order to balance the class distribution [20]. On the other side, Random Under Sampling (**RUS**) is yet another method for addressing data imbalance. In RUS, as opposed to ROS, specific samples are removed from the majority class to balance the minority class [20]. The Synthetic Minority Over-sampling Technique or **SMOTE** is another oversampling technique for imbalance data. The interpolation approach used in this technique increases the amount of new minority class instances. Before they are used to produce new minority class instances, the minority class instances that lie together are recognized. Because this technique can construct synthetic examples rather than replicating minority class instances, it avoids the problem of overfitting [21].

3.3 Model Evaluation

3.3.1 Confusion Matrix

There are numerous techniques to assess the effectiveness of learning algorithms and the classifiers they create in supervised machine learning (ML). A confusion matrix, which tracks successfully and erroneously recognized instances for each class, is used to build measures of the quality of categorization [23].

Table 4. A confusion matrix for multiclass classification

		Actual		
		Positive	Negative	Neutral
Predicted	**Positive**	*TP*	FP_1	FP_2
	Negative	FN_1	*TN*	$FNeutral_1$
	Neutral	$FNeutral_2$	FN_2	*TNeutral*

Table 4 above is the example of a confusion matrix of three class classification. *TP* (True Positive) stands for the number of comments correctly classified as positive comments, whereas *FP* (False Positive) stands for comments that are classified as positive when they are actually negative/neutral. The amount of positive comments that are classified as negative/neutral when they are actually positive is called *FN* (False Negative) or *FNeutral* (False Neutral). If a comment is predicted as negative or neutral correctly, it is considered as *TN* (True Negative) or *TNeutral* (True Neutral).

3.3.2 Precision

True positives and examples that are incorrectly categorized as positives (false positives) are two factors that affect precision [23]. The precision of a classification model is its

ability to identify only the relevant data points [24]. Calculating precision mathematically involves dividing the total number of true positives by the sum of true positives and false positives, as shown in Eq. (1).

$$precision = \frac{TP}{TP + FP} \tag{1}$$

It expresses the proportion of the data points of our model that is said existed in the relevant class.

3.3.3 Recall

Recall depends on the data that were correctly classified (true positives) and incorrectly classified (false negatives) [23]. This function assesses a model's capacity to locate every relevant examples in a data set [24]. According to the mathematical model, recall is determined by dividing the total number of true positives by the total number of true positives + the total number of false negatives, as shown in Eq. (2).

$$recall = \frac{TP}{TP + FN} \tag{2}$$

It expresses the ability to find all relevant instances of a class in a data set.

3.3.4 F1-Score

F1-score gauged the model's accuracy. The formula determines it based on accuracy and reminders [25]. F1-Score is a composite metric that rewards algorithms with better sensitivity and challenges those with higher specificity [23].

$$F1 - score = \frac{2 \times precision \times recall}{precision \times recall} \tag{3}$$

3.3.5 Accuracy

By displaying the likelihood of the class label's true value, accuracy provides an approximation of the algorithm's efficacy; in other words, it evaluates the algorithm's overall performance [23]. Accuracy is calculated as the sum of accurate predictions divided by the total number of data sets. Accuracy ranges from 0.00 to 1.0, with 1.0 being the best and 0.00 being the worst [25].

$$accuracy = \frac{TP + TN + TNeutral}{P + N + Neutral} \tag{4}$$

4 Results and Discussion

4.1 Result

The data is divided to 80% for training data and 20% for testing. Data imbalances are overcome by three methods namely SMOTE, Random Over Sampling (ROS), Random Under Sampling (RUS). Therefore, the amount of training data and testing data will

follow the results of the data imbalance processing. For the analysis, seven different classification methods are used in this study, such as CatBoost Classifier, SVM, Random Forest, Logistic Regression, GaussianNB, BernoulliNB, and Neural Network for Natural language processing (NLP) with BiGRU method. The CatBoost Classifier is tuned by the 1000 iterations parameter, 1% learning rate parameter, 42 random seeds parameter, 8 depth of the tree parameter, with accuracy and auc as loss parameter. The SVM, Random Forest, Logistic Regression, GaussianNB, and BernoulliNB are tuned by the default parameter, with 5 fold cross validation. The Bidirectional Gated Recurrent Unit (BiGRU) is created by some layers, the first layer is the embedding layer which uses a vector with dimensions of 100 words long, to represent each word. The second layer is the Bidirectional Gated Recurrent Unit (BiGRU) layer. The third layer is the Dense layer with dropout and ReLU activation function. The last layer is the Output Layer with 3 neurons and softmax as the activation function.

Table 5 below shows the comparison accuracy of each methodology with the balancing scenarios. Turns out, the best classifier methods are CatBoost Classifier with 88% accuracy and BiGRU with 83% accuracy, after the imbalance data are handled by Random Over Sampling (ROS). However, it also can be shown that the powerful classification method with the highest accuracy is the CatBoost Classifier with 75% or above in both imbalance and balanced data with ROS, RUS, and SMOTE.

Table 5. Comparison Methods Accuracy

	Imbalancd	SMOTE	ROS	RUS
CatBoost	0.75	0.76	0.88*	0.75
SVM	0.58	0.52	0.64	0.49
Random Forest	0.59	0.47	0.76	0.49
Logistic Regression	0.58	0.52	0.57	0.49
GaussianNB	0.49	0.40	0.39	0.39
BernoulliNB	0.52	0.43	0.47	0.46
BiGRU	0.56	0.58	0.83*	0.46

* indicate the high accuracy

Table 6 shows the comparison of F1-Score for each methodology with the balancing scenarios. Same as accuracy, the best classifier methods are CatBoost Classifier with 88% accuracy and BiGRU with 83% accuracy, after the imbalance data are handled by Random Over Sampling (ROS). Moreover, in Table 5 it also can be shown that the powerful classification method with the highest F1-Score is the same as the best method for the accuracy.

Table 7 shows the comparison of precision for each methodology with the balancing scenarios. The comparison of precision produces the same result as the comparison of accuracy and F1-Score. When compared by the precision, CatBoost Classifier and BiGRU, handled by ROS, still get the highest points with 89% and 84% respectively.

Table 6. Comparison Methods F1-Score

	Imbalanced	SMOTE	ROS	RUS
CatBoost	0.73	0.77	0.88*	0.75
SVM	0.49	0.51	0.65	0.50
Random Forest	0.52	0.48	0.75	0.49
Logistic Regression	0.53	0.51	0.57	0.49
GaussianNB	0.45	0.35	0.36	0.38
BernoulliNB	0.52	0.43	0.47	0.46
BiGRU	0.55	0.58	0.83*	0.46

* indicate the high F1-Score

Table 7. Comparison Methods Precision

	Imbalanced	SMOTE	ROS	RUS
CatBoost	0.81	0.79	0.89*	0.77
SVM	0.49	0.52	0.66	0.51
Random Forest	0.59	0.47	0.78	0.53
Logistic Regression	0.54	0.53	0.58	0.50
GaussianNB	0.46	0.39	0.40	0.47
BernoulliNB	0.53	0.44	0.48	0.47
BiGRU	0.55	0.58	0.84*	0.47

* indicate the high precision

Table 8 shows the comparison of Recall for each methodology with the balancing scenarios. Catboost with the recall of 88% is still the highest, followed by BiGRU with the recall of 83%. This shows that from every comparison of evaluation model, Catboost continuously produces the highest score and therefore can be said as the best classification model for this data.

Catboost and BiGRU with balance handling utilizing ROS are the best models for the full analysis employing metrics of evaluation using accuracy, precision, recall, and F1-score. The explanation for the best model, i.e. BiGRU and Catboost Classifier, will briefly show as the example of analysis process.

4.1.1 BiGRU

For the Random Over Sampling handled data, the total parameters number of BiGRU is 369.163 parameters, with 15.963 trainable parameters and 353.200 non-trainable parameters. The output will be dense to 3 neuron layers with softmax activation layers. Model summary for BiGRU is visualized by Fig. 3.

Table 8. Comparison Methods Recall

	Imbalanced	SMOTE	ROS	RUS
CatBoost	0.75	0.76	0.88*	0.75
SVM	0.58	0.52	0.64	0.49
Random Forest	0.58	0.49	0.76	0.51
Logistic Regression	0.58	0.52	0.58	0.49
GaussianNB	0.49	0.4	0.39	0.40
BernoulliNB	0.52	0.43	0.47	0.46
BiGRU	0.56	0.58	0.83*	0.46

* indicate the high recall

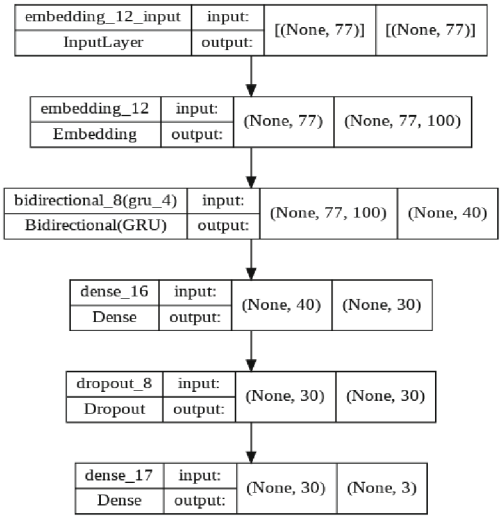
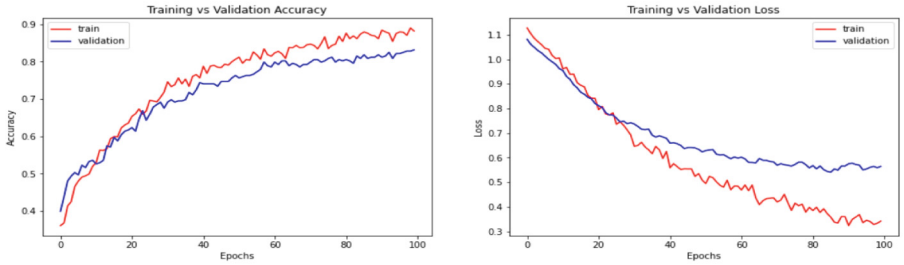


Fig. 3. Model summary of BiGRU

The BiGRU fine-tuned-based best network has a 100 epoch. In this scenario, the highest recognition rate of 89% is reached when the network is trained with Adam Optimizer, batch size of 64, and maximum epoch of 20. However, as in Fig. 4, these data show that there is underfitting in recognizing train and validation loss, so the resulting model is not very accurate.

Figure 5 shows the confusion matrix and the value of evaluation metric. After evaluating the BiGRU model, the accuracy rate of this model is 83% with an average precision level of 84%. The precision for every variable is 88% for neutral comments, 91% for negative comments, and 72% for positive comments. Negative comments is the variable with the highest precision.



(a) (b)

Fig. 4. Training vs validation plot for (a) accuracy and (b) loss

	precision	recall	f1-score	support
0	0.88	0.74	0.80	99
1	0.91	0.85	0.88	117
2	0.72	0.91	0.81	92
accuracy			0.83	308
macro avg	0.84	0.83	0.83	308
weighted avg	0.84	0.83	0.83	308

Fig. 5. Confusion matrix and the value of evaluation metric

4.1.2 Catboost

Another powerful model which has best accuracy is the CatBoost classifier with ROS. This algorithm is fine-tuned-based best network has 1000 iterations and is trained with 8 depths of tree, 1% learning rate, and 42 random seeds. The Figs. 6 and 7 show the Train and validation accuracy plot and matric of evaluation value, respectively.

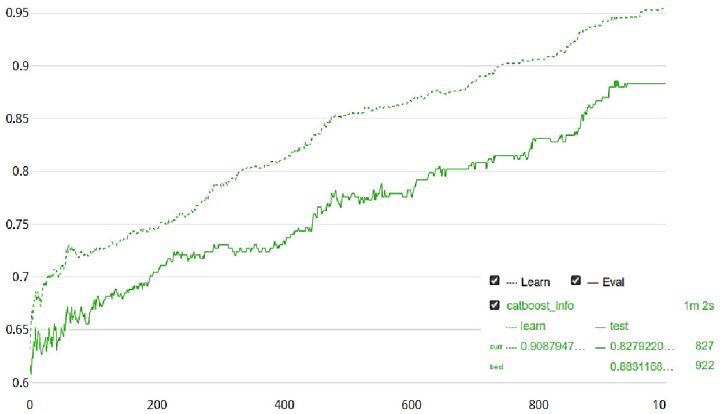


Fig. 6. Train and validation accuracy plot for Random Over Sampling

	precision	recall	f1-score	support
0	0.78	0.90	0.84	99
1	0.91	0.85	0.88	117
2	0.99	0.91	0.95	92
accuracy			0.88	308
macro avg	0.89	0.89	0.89	308
weighted avg	0.89	0.88	0.88	308

Fig. 7. Metric of Evaluation

For the Random Over Sampling handled data, the highest classification rate of 90% is reached and after evaluating the model, the accuracy rate of this model is 88% with an average precision level of 89%. The precision for every variable is 78% for neutral comments, 91% for negative comments, and 99% for positive comments. Positive comments is the variable with the highest precision.

4.2 Discussion

As discussed before in the literature review, there have been some studies about social media sentiment analysis. Hikmah, Fauzan, and Harliana [11] are using Naive Bayes to classify the sentiment from Twitter comments about Indonesian netizen perspective towards the booster vaccine, by consisting 23% of positive sentiment, 15% of neutral sentiment, and 76% of negative sentiment with the result 89% of accuracy rate. With an accuracy rate of 92.96%, Lestari and Saepudin [20] discovered tweets with positive sentiment of up to 86% and tweets with negative sentiment of up to 14% using the Naive Bayes Algorithm on sentiment analysis of the Sinovac vaccine. Ananda [22] discovered 63% of accuracy rate within Bidirectional GRU (BiGRU) methods for analyzing Indonesia's netizen sentiment to COVID-19 vaccine consisting 59% neutral sentiment, 71% negative sentiment, and 41% positive sentiment.

Based on the results of the three studies including our own, it is clear that utilizing the CatBoost Classifier to categorize sentiments is the most appropriate. It achieves 88% accuracy for overall class, 78% accuracy for neutral sentiment, 91% for negative sentiment, and 99% for positive sentiment. The data was treated with a comment using Random Over Sampling. Additionally, the use of the BiGRU method in sentiment classification can be considered, with a classification accuracy result of 83 percent accuracy for each sentiment category, namely 88 percent for neutral sentiment, 91% for negative sentiment, and 72% for positive sentiment, using the same data handling method. However, the issue of data underfitting must be revisited for the implementation of this method.

5 Conclusion and Future Research

According to our research, the Indonesian public has mixed feelings about the COVID-19 vaccine booster. After comparing 7 different methods with 4 scenarios imbalance data handling methods, the best method with the highest accuracy is CatBoost Classifier with Random Over Sampling method. The result of the analysis showed 78% for

neutral comments, 91% for negative comments, and 99% for positive comments. After preprocessing the data by categorizing the data, cleaning the data from typo, tokenizing the data, embedding the data, handling the imbalance data, the BiGRU shows there's some problem with the fitting data (underfitting).

Mining data from social media sometimes required permit that could take sometimes to acquire. Due to many vernaculars and dialect in Indonesia, many comments contained or were fully in other dialect which complicate the cleaning process. Further researches with the same type of data are expected to use more data and more cleaning of the data collection in an attempt to fix the fitting problem.

Acknowledgement. The authors thanks to the Department of Statistics and Research Center for Artificial Intelligence and Big Data, Universitas Padjadjaran which supports this research. The work of Tutut Herawan and Eka Novita Sari is supported by AMCS Research Center.

References

1. WHO Coronavirus (COVID-19) Dashboard, https://covid-19.who.int retrieved 16 June 2022
2. Jeyanathan, M., Afkhami, S., Smail, F., Miller, M.S., Lichty, B.D., Xing, Z.: Immunological considerations for COVID-19 vaccine strategies. Nature Reviews Immunology **20**, 615–632 (2020)
3. Lurie, N., Saville, M., Hatchett, R., Halton, J.: Developing Covid-19 vaccines at pandemic speed. N. Engl. J. Med. **382**, 1969–1973 (2020)
4. Jokowi Bentuk timnas Percepatan Pengembangan Vaksin Covid-19, Jokowi Bentuk Timnas Percepatan Pengembangan Vaksin Covid-19 | Republika Online Retrieved 18 June 2022
5. What You Need To Know About COVID-19 Vaccines, What you need to know about COVID-19 vaccines | UNICEF Indonesia Retrieved 18 June 2022
6. Program Vaksinasi COVID-19 Mulai Dilakukan Presiden Orang Pertama Penerima Suntikan Vaksin COVID-19, http://p2p.kemkes.go.id/program-vaksinasi-covid-19-mulai-dilakukan-presiden-orang-pertama-penerima-suntikan-vaksin-covid-19/ Retrieved 18 June 2022
7. Coronavirus (Covid-19) Vaccinations, Coronavirus (COVID-19) Vaccinations - Our World in Data Retrieved 18 June 2022
8. Vaksinasi COVID-19 Nasional, Vaksin Dashboard (kemkes.go.id) Retrieved 18 June 2022
9. Interim Statement on Booster Dose for COVID-19 Vaccination, Interim statement on booster doses for COVID-19 vaccination (who.int) Retrieved 18 June 2022
10. Shekhar, R., Garg, I., Pal, S., Kottewar, S., Sheikh, A.B.: COVID-19 vaccine booster: to boost or not to boost. Infectious disease reports **13**(4), 924–929 (2021)
11. Mizrahi, B., et al.: Correlation of SARS-CoV-2-breakthrough infections to time-from-vaccine. Preliminary Study (2021)
12. Zhan, Y., Yin, H., Yin, J.Y.B.: 1.617. 2 (Delta) Variant of SARS-CoV-2: features, transmission and potential strategies. International journal of biological sciences **18**(5), 1844 (2022)
13. Vaksin Booster Bertujuan Melindungi Dan Menjadi Modal Pemulihan Ekonomi, Vaksin Booster Bertujuan Melindungi Dan Menjadi Modal Pemulihan Ekonomi | Covid19.go.id Retrieved 18 June 2022
14. Ong, S., Pauzi, M.B.M., Gan, K.H.: Text Mining and Determinants of Sentiments Towards the COVID-19 Vaccine Booster of Twitter Users in Malaysia. MDPI, Basel, Switzerland (2022)
15. Yue, L., Chen, W., Li, X., Zuo, W., Yin, M.: A survey of sentiment analysis in social media. Knowl. Inf. Syst. **60**(2), 617–663 (2018). https://doi.org/10.1007/s10115-018-1236-4

16. Hikmah, K., Fauzan, A.C., Harliana. Sentiment Analysis of Vaccine Booster during Covid-19: Indonesian Netizen Perspective Based on Twitter Dataset. Teknologi Komputer dan Sistem Informasi, STMIK Pringsewu (2022)
17. Balahur, A.: Sentiment Analysis in Social Media Texts, pp. 120–128 (2013)
18. Narasamma, V.L., Sreedevi, M.: Twitter based Data Analysis in Natural Language Processing using a Novel Catboost Recurrent Neural Framework. European Commission Joint Research Centre (2021)
19. Pristiyono, R., Ihsan, M., Anjar, M.A.A.A., Rambe, F.H.: Sentiment Analysis of COVID-19 Vaccine in Indonesia Using Naive Bayes Algorithm IOP Conf. Ser.: Mater. Sci. Eng. **1088**, 012045
20. Ananda, D.: Analisis Sentimen Pengguna Media Sosial Terhadap Kebijakan Pemerintah Dalam Program Vaksinasi Covid-19 Di Indonesia Menggunakan Metode Bidirectional Gated Recurrent Unit (BiGRU). Bachelor thesis. Department Of Statistics, Universitas Padjadjaran, Sumedang (2022)
21. Jeatrakul, P., Wong, K.W., Fung, C.C.: "Classification of Imbalanced Data by Combining the Complementary Neural Network and SMOTE Algorithm" School of Information Technology. Murdoch University, Australia (2010)
22. Sabarmathi, G., Chinnaiyan, R.: Sentiment Analysis for Evaluating the Patient Medicine Satisfaction. Int. J. Computat. Intell. Cont. **13**, 113–118 (2021)
23. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond Accuracy, F-Score and ROC: a Family of Discriminant Measures for Performance Evaluation. In: Sattar, A., Kang, Bh. (eds.) AI 2006: Advances in Artificial Intelligence. Lecture Notes in Computer Science(), vol 4304. Springer, Berlin, Heidelberg (2006)
24. When Accuracy Isn't Enough, Use Precision and Recall to Evaluate Your Classification Model, When Accuracy Isn't Enough, Use Precision and Recall to Evaluate Your Classification Model Retrieved 22 June 2022
25. Vujovic, Z.D.: Classification Model Evaluation Metrics. Int. J. Adv. Comput. Sci. Appl. **12**(6), 1–8 (2021)