Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Research Article

Improving Multi-label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation Using IndoBERT Model

Leno Dwi Cahya [a,*], *Ardytha Luthfiarta* [b] , *Julius Imanuel Theo Krisna* [c] , *Sri Winarno* [d] , *Adhitya Nugraha* [e]

[a,b,c,d,e] *Universitas Dian Nuswantoro, Jl. Imam Bonjol No.207, Pendrikan Kidul, Kec. Semarang Tengah, Kota Semarang, Jawa Tengah 50131, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 14 November 2023
Revisi Akhir: 12 Januari 2024
Diterbitkan *Online*: 16 Januari 2024

KATA KUNCI

Imbalanced,
SMOTE,
Augmentation,
Sentiment,
NLP

KORESPONDENSI

E-mail: lenodwicahya@gmail.com*

ABSTRACT

Sentiment and emotion analysis is a common classification task aimed at enhancing the benefit and comfort of consumers of a product. However, the data obtained often lacks balance between each class or aspect to be analyzed, commonly known as an imbalanced dataset. Imbalanced datasets are frequently challenging in machine learning tasks, particularly text datasets. Our research tackles imbalanced datasets using two techniques, namely SMOTE and Augmentation. In the SMOTE technique, text datasets need to undergo numerical representation using TF-IDF. The classification model employed is the IndoBERT model. Both oversampling techniques can address data imbalance by generating synthetic and new data. The newly created dataset enhances the classification model's performance. With the Augmentation technique, the classification model's performance improves by up to 20%, with accuracy reaching 78%, precision at 85%, recall at 82%, and an F1-score of 83%. On the other hand, using the SMOTE technique, the evaluation results achieve the best values between the two techniques, enhancing the model's accuracy to a high 82% with precision at 87%, recall at 85%, and an F1-score of 86%.

1. INTRODUCTION

Many companies make various efforts to increase sales based on consumer comfort by conducting sentiment analysis derived from customer opinions or comments to enhance the quality of their products and services. Along with the rapid development of the internet and social media, obtaining significant data sources regarding consumer opinions is easy. Companies can quickly gather consumer opinions through social media platforms such as Twitter, Facebook, and Instagram[1]. Sentiment analysis has recently been widely accepted among researchers and the business world, governments, and organizations [2]. Along with technological advancements, sentiment analysis has started incorporating Artificial Intelligence (AI) and Natural Language Processing (NLP). Sentiment analysis under machine learning is part of natural language processing or NLP.

However, some sentiment or consumer opinion data limitations make these machine-learning algorithms less effective. One common and critical challenge experienced is the issue of imbalanced datasets. It occurs when the class distribution is uneven, with a significant difference in the number of instances. With this problem, the classification model tends to be biased toward the dominant class, potentially causing data with less significant characteristics (minorities) to be misclassified into the majority class[3]. This results in the testing of the model on classes with fewer instances being unbalanced, affecting overall accuracy or causing a decrease in model accuracy.

In several previous studies related to addressing imbalances in image or picture data, there are various solutions for class

https://doi.org/10.25077/TEKNOSI.v9i3.2023.290-298

imbalances, including assigning weights to minority classes, using sampling techniques, determining similarity loss functions, and using focal loss [4]. One method for addressing imbalanced datasets is by employing sampling techniques, namely, oversampling and undersampling. Through oversampling, the minority class data is duplicated to match the number of instances in the majority class. In contrast, undersampling involves removing cases from the majority class until its data size matches that of the minority class. With oversampling, the resulting dataset balances minority and majority class data, although there is a risk of the model overfitting.

On the other hand, undersampling may result in removing important information from the majority class, leading to decreased accuracy when classifying majority class data [3]. Based on previous research, oversampling techniques are more commonly used as a resolution method due to their better outcomes than undersampling methods. The oversampling technique has been further developed by introducing synthetic data instead of duplicating the dataset. This technique is known as the Synthetic Minority Over-sampling Technique (SMOTE). The SMOTE technique enhances the model's classification ability by adding synthetic data to the previous dataset, generated through feature-based approaches using the k-nearest neighbor (KNN) calculation [5], [6].

SMOTE is a method used on numerical data requiring numerical representation or feature extraction. This process will display values of features represented as a list of words. Two feature extraction approaches are used, namely Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF has better capabilities than BoW because TF-IDF assigns weight values to the importance of words, whereas BoW only displays the word count. In another study [7], it is shown that a classification model with imbalanced data handling using SMOTE with TF-IDF can increase the cardinality of minority class labels through significant improvements in Macro Avg Recall and Macro Avg F1-Score compared to the baseline model. [8].

Augmentation is one of the techniques used to generate data with new variations. In text processing, augmentation is performed by adding, deleting, or changing words in the existing data [9]. Word embedding can be used to modify or introduce new words near the original words and those present in the word embedding. Through word embedding, new sentences that semantically carry the same meaning or synonym as the previous dataset are generated but with variations in different words.

Many machine learning models for Natural Language Processing (NLP) have been developed for sentiment analysis with imbalanced data. These include conventional machine learning models like Naïve Bayes Classifier (NBC), Support Vector Machine (SVM), and deep learning models based on Neural Networks. Examples of these deep learning models include Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), and even the latest models based on Transformers, such as Bidirectional Encoder Representations from Transformers (BERT) [10]–[12].

In studies on text classification in the Indonesian language, the IndoBERT model has demonstrated optimal capabilities compared to other models in text classification. In 2023, Juarto and Yulianto researched text classification using several language models, including IndoBERT, XLMNet, XLM Roberta, and MultilingualBERT [13]. In this study, IndoBERT emerged as the model with the highest accuracy compared to the other models. Similarly, in research on sentiment classification conducted by Saadah et al. in 2022the best-performing model was found to be IndoBERT compared to IndoBERTweet and CNN-LSTM [12].

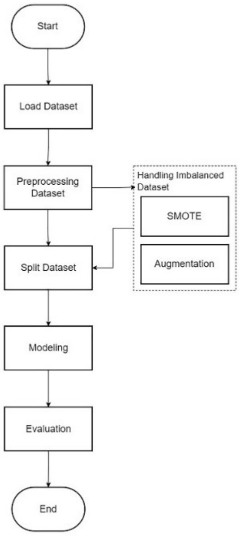


Figure 1. Research method

A study conducted by Ningsih et al. [14] on word embedding, specifically replacing vocabulary with synonyms, demonstrated an impact on the testing of classification models. In another study [15], the word embedding technique involved adding words from

https://doi.org/10.25077/TEKNOSI.v9i3.2023.290-298

pre-training or pre-trained data to the BERT model. The results consistently showed superiority compared to using Word2Vec and WordNet techniques. In Indonesian language datasets, IndoBERT has proven to enhance model performance [16].

This research investigates the impact of using newly augmented and SMOTE-generated data on imbalanced text datasets in the Indonesian language. It is hoped that the results of this study will contribute to the development of more accurate and effective classification methods in addressing dataset imbalances, particularly in natural language processing.

2. METHOD

This research is conducted through several stages. As depicted in Figure 1, the study begins with data retrieval from a labeled or supervised text dataset. It is then followed by pre-processing to clean the data from unnecessary characters. Subsequently, data imbalance is addressed using SMOTE and Augmentation. The balanced data is utilized in the classification model, and the model results are evaluated for both dataset handling methods.

2.1. Dataset

In a study conducted by [17], a multi-label, multi-class dataset in the Indonesian language was created based on public reviews of mobile applications with sentiment and emotional values. Another supporting factor for developing this dataset is the limited availability of textual datasets in the Indonesian language that are based on multi-label multi-class for sentiment analysis tasks, especially those related to text classification. The data generated by this research was cleaned and processed during the pre-processing steps and annotated with three sentiments (Positive, Negative, and Neutral) and six emotions (Anger, Fear, Sadness, Happiness, Love, and Neutral). The total number of datasets produced from the set of comments is 21,694.

Table 1. Dataset samples

Content	Sentiment	Emotion
Sangat membantu sekali bagi yang ingin tau bahasa yang benar.	Positive	Happy
Saya Suka Dengan Aplikasi Ini Karena Saya Kepengin Punya Pacar	Positive	Love
Tolong di perbaiki lagi sering keluar sendiri dan sering lag	Negative	Sad
Ga ada tombol log out atau ganti akun, kok aneh!?	Negative	Fear
aplikasi apa ini masak chat harus bayar ga mutu bnget	Negative	Anger
Semoga aja kalo update nanti hujan nya seperti asli	Neutral	Neutral

There are two labels with imbalanced data distribution in each class. The sentiment label has a relatively balanced distribution with a not-too-significant difference in the amount of data. However, the emotion class distribution exhibits a significant imbalance, with a considerable difference in the number of data for each class.

2.2. Pre-processing

The dataset [17] from the application reviews has undergone pre-processing, including (1) removing URLs, (2) removing mentions, hashtags, and special characters, (3) removing emojis, (4) removing duplicates, and (5) removing line breaks or new line formats. In this study, additional pre-processing is conducted due to sentences containing many irrelevant characters or punctuation marks. Text case folding is applied, converting the text to lowercase. Additionally, data that lacks meaning, such as data consisting of only a few letters, is removed. Furthermore, pre-processing results without value or missing values are also eliminated.

2.3. Handling Imbalanced Datasets

This study employed oversampling methods using two approaches, namely utilizing the Synthetic Minority Over-sampling Technique (SMOTE) and conducting dataset augmentation.

2.3.1. SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is used to handle imbalanced data. SMOTE is an oversampling method that randomly generates samples and is based on the concept of the nearest neighbors. SMOTE is typically used on numerical data, so for application in text datasets, vectorization is required to produce a numerical representation of each word, which forms a set of features. The research conducted by [8] indicates that feature extraction using the TF-IDF technique within SMOTE produces better classification accuracy than Bag of Words (BoW). TF-IDF records the importance of a word, whereas BoW only counts the occurrences of a word. Therefore, in this study, the TF-IDF method is employed to represent the weight of each word, calculated by multiplying Term Frequency (TF) and Inverse Document Frequency (IDF) as shown in equation (1).

$$tf \cdot idf_{t,d} = tf_{t,d} \times idf_{t,d} \qquad (1)$$

$$tf \cdot idf_{t,d} = \frac{count\ of\ term\ t\ in\ document\ d}{total\ number\ of\ terms\ in\ document\ d} \times log\left(\frac{N}{df_t}\right) \qquad (2)$$

Term Frequency is a method used to calculate the occurrence of words in each document, providing a higher weight to frequently appearing words. Inverse Document Frequency calculates the weighting of words across the entire text corpus, so words with fewer occurrences across the corpus will have a higher weight. By combining both TF and IDF values, the TF-IDF assigns weights to each word in a document based on how often the word appears in the document (TF) and how common the word is across the entire corpus (IDF). This helps identify unique and important words, as shown in equation (2).

2.3.2. Text Data Augmentation

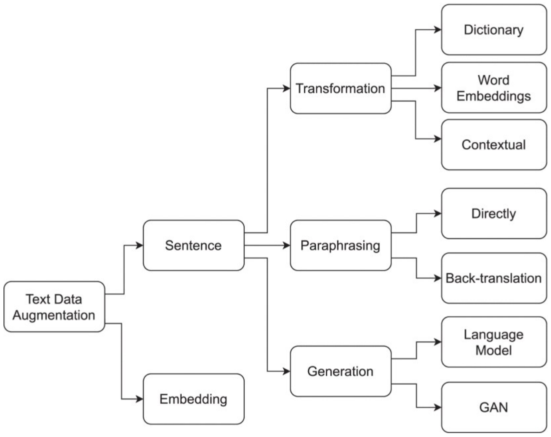


Figure 2. Taxonomy of text data augmentation methods [18]

Several methodological approaches can be used for text augmentation. In sentence composition, words can be augmented using various approaches, including the following [18]:

a. Transformation is a simple lexical operation approach performed on words in a sentence to create variations in the original sentence, commonly known as EDA or Easy Data Augmentation. EDA is a widely used method that involves synonym replacement, word swapping, word insertion, and word deletion [9]. Synonym replacement is one of the most common practices in this subcategory. Some strategies also utilize language models and pre-trained word embedding models to replace words with similar ones [15].
b. Paraphrasing, using paraphrasing techniques to obtain sentence variations reformulated from the original samples. One commonly used method is back-translation (BT), which involves translating the original sentence into an intermediate language and then translating it back into the original language [19].
c. This strategy focuses more on generating entirely new samples based on the original dataset. This method differs from previous approaches as it modifies words or original sentences and creates samples using language models, such as GPT-2 [3].

2.4. Split Data

After generating datasets with three scenarios, namely without handling data imbalance and after addressing data imbalance with SMOTE and Augmentation. Before being used for modeling, the resulting datasets undergo data splitting into training data, validation data, and test data. The dataset is divided with a ratio of 80% for training data, 10% for validation data, and 10% for test data, maintaining the same class distribution in each dataset.

2.5. Model

The classification is performed using the IndoBERT model, which stands for Indonesia Bidirectional Encoder Representations Transformers. IndoBERT is a BERT-based mode [20] designed for the Indonesian language and trained using the Transformer architecture. In the Transformer architecture, an encoder-decoder architecture is initially designed for machine translation, but in BERT, only the Encoder architecture is used.

BERT utilizes the encoder architecture from Transformers for contextual understanding in the language. There are two main stages in the BERT model, namely Pre-training and Fine-tuning.

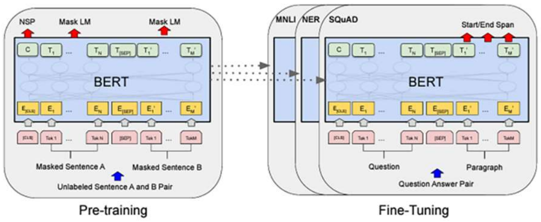


Figure 3. BERT architecture [20]

The model is trained unsupervised on a huge dataset in the pre-training stage. For the IndoBERT model, a large-sized dataset was collected, consisting of around four billion words from Indonesian text data. This dataset includes text from sources such as local online news, social media, Wikipedia, online articles, subtitle texts from video recordings, and a parallel dataset called Indo4B. Indo4B covers both formal and casual sentences in the Indonesian language [21]. The pre-training process has several main stages, namely the Masked Language Model (MLM) and Next Sentence Prediction (NSP). Before the MLM process, Tokenization is performed to represent the text as vector embeddings with additional tokens, namely [CLS] at the beginning and [SEP] at the end of each sentence. In the MLM stage, a certain number of tokens are replaced with [MASK] to be predicted correctly by the model.

Meanwhile, in NSP, predictions are made about the relationship between sentences whether they are connected or not. In the fine-tuning stage, the trained model can be used for specific tasks with the same architecture as the pre-trained model but with a smaller dataset. Adjustments to the parameters used are also necessary to support the specific tasks being performed.

2.6. Evaluation

Evaluation is used to measure the model's classification performance on the generated datasets using a confusion matrix by calculating Accuracy, Precision, Recall, And F1-Score. The confusion matrix employs the concepts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Accuracy indicates the ratio of correct predictions to the total predictions made by the classifier on the test data.

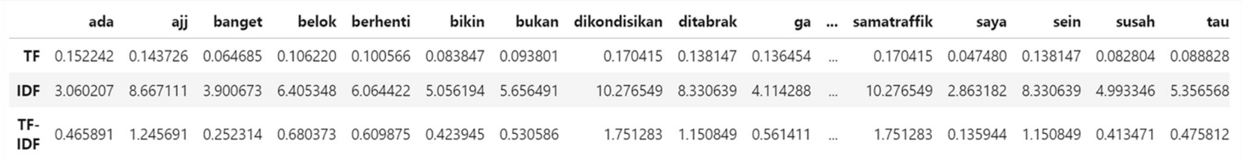
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision, also known as positive predictive value, represents the true positive value among all positive predictions.

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

Recall, often called sensitivity, represents the relative number of correctly classified positive instances among all positive instances.

https://doi.org/10.25077/TEKNOSI.v9i3.2023.290-298



	ada	ajj	banget	belok	berhenti	bikin	bukan	dikondisikan	ditabrak	ga	...	samatraffik	saya	sein	susah	tau
TF	0.152242	0.143726	0.064685	0.106220	0.100566	0.083847	0.093801	0.170415	0.138147	0.136454	...	0.170415	0.047480	0.138147	0.082804	0.088828
IDF	3.060207	8.667111	3.900673	6.405348	6.064422	5.056194	5.656491	10.276549	8.330639	4.114288	...	10.276549	2.863182	8.330639	4.993346	5.356568
TF-IDF	0.465891	1.245691	0.252314	0.680373	0.609875	0.423945	0.530586	1.751283	1.150849	0.561411	...	1.751283	0.135944	1.150849	0.413471	0.475812

Figure 6. TF-IDF numerical representation in a document

$$Recall = \frac{TP}{TP+F} \tag{5}$$

Based on the precision and recall values, the harmonic mean yields the F1-Score. The F1-Score provides a balance between precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

3. RESULT

3.1. *Dataset*

The study conducted by Riccosan and Saputra in 2023 resulted in a multi-label dataset with sentiment and emotion labels obtained from application reviews in the Indonesian language. However, the connections between each class are not evenly distributed. The negative class is only connected to sadness, anger, and fear, while the positive class is only connected to happiness and love. The Neutral class in sentiment is only connected to the Neutral class in emotion. Therefore, by knowing the emotion, the sentiment can be determined. Hence, the imbalanced data handling in this study focuses on emotion labels, namely Sad, Anger, Fear, Happy, Love, and Neutral.

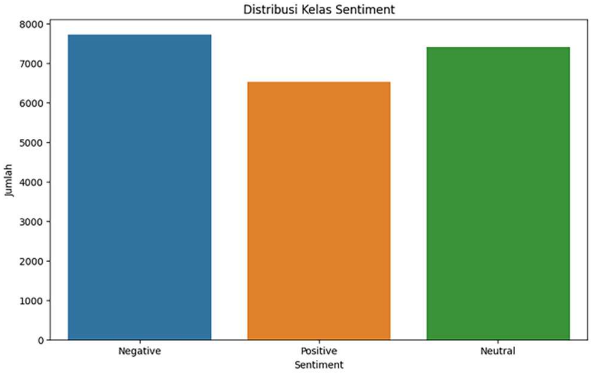


Figure 4. Bar plot of sentiment class distribution

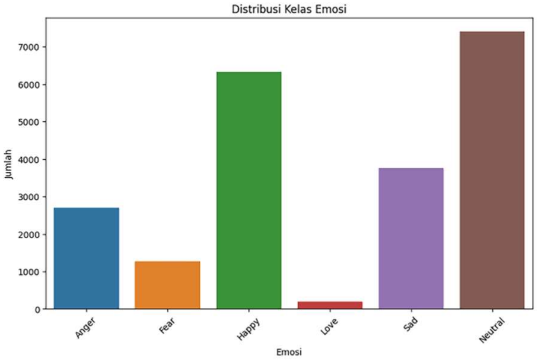


Figure 5. Bar plot of emotion class distribution

Before handling the imbalanced dataset, pre-processing was performed on the dataset. In the previous study, the dataset had undergone pre-processing. Additional pre-processing was conducted by applying text case-folding, transforming the text into lowercase. Additionally, characters or punctuation marks without meaning or in excess were removed. As a result, there were data consisting of only a few irrelevant letters with no value or missing value, necessitating cleaning.

Table 2. Class distribution before and after pre-processing

Sentiment	Emotion	Total	
		Before	**After**
Negative	Sad	3,753	3,748
	Anger	2,697	2,675
	Fear	1,271	1,269
Positive	Happy	6,330	6,182
	Love	193	190
Neutral	Neutral	7,453	7,307

3.2. *Imbalanced Handling*

3.2.1. SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is a method used with numerical data, so it is necessary to perform a numerical representation of the features in the dataset, also known as feature extraction from data in terms or tokens. Commonly used numerical representations are Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). In this study, the TF-IDF method is employed, assigning weights to each word in a document based on how often that word appears in the document (TF) and how common the word is across the entire corpus (IDF)

https://doi.org/10.25077/TEKNOSI.v9i3.2023.290-298

Through numerical representation using TF-IDF, as shown in Figure 6, text data can be oversampled using SMOTE. The SMOTE method randomly selects minority samples to obtain *k* nearest neighbors from that sample. New samples can be created from these nearest neighbors, resulting in many new samples equal to most data [5]. As shown in Table 3, the new data obtained has variations in similar words, but the word order becomes irregular, causing the semantics of the data to be lost.

Table 3. Example of data generated by SMOTE

Original Sentence	*aku sangat suka aplikasi ini*
SMOTE	*aplikasi dengan sangat suka* *app dengan sangat suka ini saya* *mksh app dg sangat suka ini*

3.2.2. *Text Data Augmentation*

Many approaches can be taken for text augmentation, such as transformation, paraphrasing, and generation [18]. This study applied augmentation approaches to transformation and paraphrasing due to limited resources for Bahasa Indonesia in a generation.

The transformation technique in augmentation uses the EDA or Easy Data Augmentation method. EDA [9] is a simple word replacement-based augmentation method that involves adding, swapping, and deleting words. One commonly used technique is to replace words with synonyms. This study used WordNet in the Indonesian language to find synonyms for words in the dataset. In addition to using synonyms, augmentation was performed by contextually adding text. Contextual Augmentation leverages the capabilities of a pre-trained model to generate text variations that are not only semantically similar but also contextually similar [16]. This study's contextual augmentation model is the Indonesian BERT or IndoBERT.

The paraphrasing approach using Backtranslation. Back translation is an augmentation method that creates new words by translating the original text into an intermediate language and then back into the original language. The model used for backtranslation in this study is trained from [21], translating text from Indonesian to English (opus-mt-id-en) and translating it back to Indonesian (opus-mt-en-id).

Table 4. Example of augmented data.

Original Sentence	*aku sangat suka aplikasi ini*
Sinonym	*saya sangat berkenan dengan app ini*
ContextualWordEmbs	*saya putar suka segera app ini*
BackTranslation	*Aku sangat suka aplikasi ini*

After handling the imbalanced dataset, the distribution of each class is now balanced, as seen in Figure 7. Each class of emotion labels has a count of 7,307.

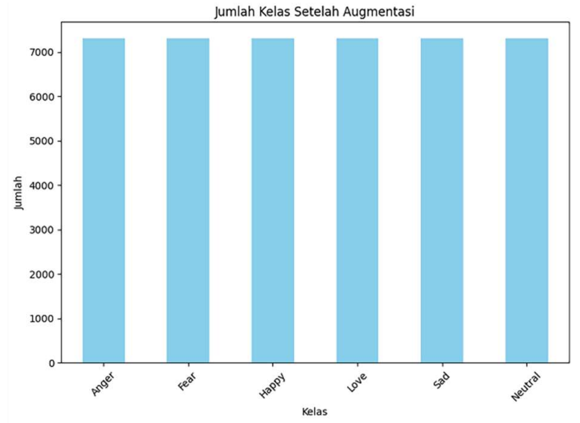
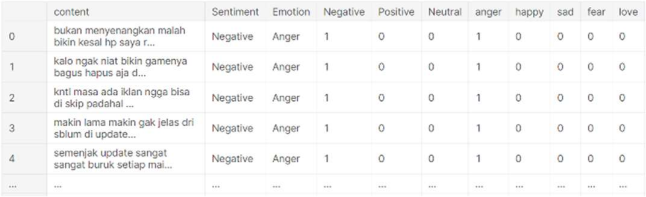


Figure 7. Boxplot of class distribution after handling imbalanced data

Label encoding was performed into multi-label after obtaining a balanced data distribution, as shown in Figure 8, for use in the classification model.



	content	Sentiment	Emotion	Negative	Positive	Neutral	anger	happy	sad	fear	love
0	bukan menyenangkan malah bikin kesal hp saya r...	Negative	Anger	1	0	0	1	0	0	0	0
1	kalo ngak niat bikin gamenya bagus hapus aja d...	Negative	Anger	1	0	0	1	0	0	0	0
2	kntl masa ada iklan ngga bisa di skip padahal ...	Negative	Anger	1	0	0	1	0	0	0	0
3	makin lama makin gak jelas dri sblum di update...	Negative	Anger	1	0	0	1	0	0	0	0
4	semenjak update sangat sangat buruk setiap mai...	Negative	Anger	1	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...

Figure 8. Label encoding

3.3. *Fine-Tuning Model*

To use the pre-trained model from IndoBERT, it is necessary to encode the text from the dataset into tensor format. This can be done through Tokenization using BertTokenizer. The pre-trained model used in the tokenizer and classification model is 'indobenchmark/indobert-base-p1' with parameters of 124.5M and previous training data from Indo4B. This study uses the AdamW optimizer with a learning rate of 1e-5 is used. The batch size used in this training is 16, with ten iterations (epochs).

3.4. *Evaluation*

Based on the average evaluation results for each class in Table 5, it can be observed that before addressing imbalanced data, the model's performance for classification achieved an accuracy of 65% with a precision of 66%, recall of 60%, and an F1-Score of 62%. After handling imbalanced data using augmentation, the classification model's performance improved by 20%, with an accuracy of 78%, precision of 85%, recall of 81%, and an F1-score of 83%. The same improvement is observed in the results obtained from SMOTE. The evaluation results significantly enhanced the model's performance, with an accuracy of 82%, precision of 87%, recall of 85%, and an F1-score of 86%. Thus, handling imbalanced data using SMOTE yielded the highest results in this study.

Table 5. Evaluation metrics

Dataset	*Precision*	*Recall*	*F1-score*	*Accuracy*
Data Asli	0.66	0.60	0.62	0,65
SMOTE	0.87	0.85	0.86	0,82
Augmentasi	0.85	0.82	0.83	0,78

4. DISCUSSION

Based on Figure 7, oversampling methods with two approaches, namely SMOTE and Augmentation, can generate synthetic or new data that can provide an equal number for each class. The numerical representation of the dataset using TF-IDF can be used to create synthetic data using SMOTE by considering the word occurrence and uniqueness. Handling imbalanced data using SMOTE produces a dataset with unchanged features from the original dataset. Based on Table 3, which is an example of the SMOTE results from one of the data, new data can be generated with a combination of words obtained through the neighborhood of TF-IDF values.

In Table 5, the evaluation matrix shows that the dataset resulting from SMOTE can provide performance in the classification model with the highest values compared to the data from augmentation. In research conducted by Abonizio et al., some existing augmentation techniques are still limited and cannot address text issues significantly [18]. In addition, based on the generated dataset, SMOTE data comes from a combination of existing features or vocabulary, making it easier for the model to process and understand similar data. However, the dataset resulting from SMOTE eliminates the semantics or meaning of the data.

As shown in Table 4, Augmentation results display newly generated datasets. It can provide datasets with the same semantics as the original data. In SMOTE, the resulting dataset undergoes minimal changes, only a rearrangement of words or changes in words with proximity. Meanwhile, the augmented dataset undergoes changes from old to new words with contextual word embedding methods that introduce new words derived from embeddings in IndoBERT. Additionally, in augmentation using synonyms and back translation, new data with the same meaning as the original data is added.

5. CONCLUSION

Based on the results of handling imbalanced datasets through oversampling methods using synthetic data or segmentation, there is an improvement in model performance. The model's ability in classification for each class can be more balanced and experience an increase. Before addressing the imbalanced dataset with the IndoBERT model, it could only achieve an accuracy of 65% with a precision of 66%, recall of 60%, and an F1-score of 62%, which was low for the minority dataset. After addressing the imbalanced dataset with various oversampling techniques, the minority class has an equal number of samples as the majority class. After balancing the dataset, the model's performance improves. Using various data augmentation methods increases the model accuracy to 78% with a precision of 85%, recall of 82%, and an F1-score of 83%. Meanwhile, the dataset resulting from SMOTE shows a higher accuracy of 0.82 with a precision of 87%, recall of 85%, and an F1-score of 86%.

There are some differences in the datasets generated by both methods. The augmented dataset experiences the addition of data in each minority class by introducing new vocabulary and adding more features or tokens, resulting in a larger dimension. It can make the model struggle to find relevant patterns in the data, leading to longer computational times to generate new data. In contrast, SMOTE does not introduce new features, making it easier for the model to understand existing patterns. However, the SMOTE-generated dataset produces more duplicate data due to limited features, while augmentation produces fewer duplicate data.

The choice between the two approaches depends on the characteristics of the text data used. If information and text structure are crucial, data augmentation may be considered to preserve the integrity of the original text. However, if the vocabulary needs are already met and precise numerical representation is essential, SMOTE might be more appropriate. Introducing new data can train the model to recognize more diverse texts or documents. Many augmentation methods can be used to generate new data, but resources for text augmentation in Indonesia are limited and need further development.

BIBLIOGRAPHY

[1] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, p. 100003, Mar. 2023, doi: 10.1016/j.nlp.2022.100003.

[2] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," *IEEE Trans Knowl Data Eng*, vol. 35, no. 11, pp. 11019–11038, Nov. 2023, doi: 10.1109/TKDE.2022.3230975.

[3] E. Alemayehu and Y. Fang, "A Submodular Optimization Framework for Imbalanced Text Classification With Data Augmentation," *IEEE Access*, vol. 11, pp. 41680–41696, 2023, doi: 10.1109/ACCESS.2023.3267669.

[4] A. Nugroho, M. A. Soeleman, R. Anggi Pramunendar, A. Affandy, and A. Nurhindarto, “Peningkatan Performa Ensemble Learning pada Segmentasi Semantik Gambar dengan Teknik Oversampling untuk Class Imbalance,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 4, pp. 899–908, 2023, doi: 10.25126/jtiik.2023106831.

https://doi.org/10.25077/TEKNOSI.v9i3.2023.290-298

[5] Z. Hengyu, "Improved SMOTE algorithm for imbalanced dataset," in *2020 Chinese Automation Congress (CAC)*, IEEE, Nov. 2020, pp. 693–697. doi: 10.1109/CAC51589.2020.9326603.

[6] B. Jonathan, P. H. Putra, and Y. Ruldeviyani, "Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, IEEE, Jul. 2020, pp. 81–85. doi: 10.1109/IAICT50021.2020.9172033.

[7] M. S. N. M. Danuri, R. A. Rahman, I. Mohamed, and A. Amin, "The Improvement of Stress Level Detection in Twitter: Imbalance Classification Using SMOTE," in *2022 IEEE International Conference on Computing (ICOCO)*, IEEE, Nov. 2022, pp. 294–298. doi: 10.1109/ICOCO56118.2022.10031684.

[8] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.

[9] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.11196

[10] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif Intell Rev*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[11] Y. Yanfi, Y. Heryadi, L. Lukas, W. Suparta, and Y. Arifin, "Sentiment Analysis of User Review on Indonesian Food and Beverage Group using Machine Learning Techniques," in *2022 IEEE Creative Communication and Innovative Technology (ICCIT)*, IEEE, Nov. 2022, pp. 1–5. doi: 10.1109/ICCIT55355.2022.10118707.

[12] S. Saadah, Kaenova Mahendra Auditama, Ananda Affan Fattahila, Fendi Irfan Amorokhman, Annisa Aditsania, and Aniq Atiqi Rohmawati, "Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 648–655, Aug. 2022, doi: 10.29207/resti.v6i4.4215.

[13] B. Juarto and Yulianto, "Indonesian News Classification Using IndoBert," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 454–460, 2023.

[14] F. S. S. Ningsih *et al.*, "Synonym-based Text Generation in Restructuring Imbalanced Dataset for Deep Learning Models," in *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*, IEEE, Mar. 2022, pp. 1–6. doi: 10.1109/NISS55057.2022.10085156.

[15] L. Hu, C. Li, W. Wang, B. Pang, and Y. Shang, "Performance Evaluation of Text Augmentation Methods with BERT on Small-sized, Imbalanced Datasets," in *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*, IEEE, Dec. 2022, pp. 125–133. doi: 10.1109/CogMI56440.2022.00027.

[16] F. Muftie and M. Haris, "IndoBERT Based Data Augmentation for Indonesian Text Classification," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*, IEEE, Aug. 2023, pp. 128–132. doi: 10.1109/ICITRI59340.2023.10250061.

[17] Riccosan and K. E. Saputra, “Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review,” *Data Brief*, vol. 50, p. 109576, Oct. 2023, doi: 10.1016/j.dib.2023.109576.

[18] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 657–668, Oct. 2022, doi: 10.1109/TAI.2021.3114390.

[19] D. R. Beddiar, M. S. Jahan, and M. Oussalah, "Data expansion using back translation and paraphrasing for hate speech detection," *Online Soc Netw Media*, vol. 24, p. 100153, Jul. 2021, doi: 10.1016/j.osnem.2021.100153.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[21] J. Tiedemann and S. Thottingal, "OPUS-MT-Building open translation services for the World," 2020. [Online]. Available: http://opus.nlpl.eu

NOMENCLATURE

$tf_{t,d}$	frequency of terms t in a document d
idf_t	invers document frequency
N	total number of documents
df_t	number of documents containing the term t
TP	positive class that was successfully predicted correctly
FP	positive class that failed to predict correctly
TN	negative class that was successfully predicted correctly
FN	negative class that failed to predict correctly

AUTHOR'S BIOGRAPGHY

Leno Dwi Cahya
An undergraduate student in Computer Science at Dian Nuswantoro University (UDINUS). a research assistant in "Bengkel Koding", a program under the Department of Computer Science UDINUS. my research focus is on machine learning and natural language processing (NLP)

Ardytha Luthfiarta, M.Kom,
Currently working as a Lecturer at Informatics Engineering Department, Computer Science Faculty, Universitas Dian Nuswantoro. He was graduated from Master of Software Engineering and Intelligent System, University Teknikal

Malaysia Malacca. He developed a passion for Research and Education in Artificial Intelligence, Data Mining, Natural Language Processing, and Deep Learning.

Julius Imanuel Theo Krisna
An undergraduate student in Computer Science at Dian Nuswantoro University (UDINUS). A research assistant in "Bengkel Koding", a program under the Department of Computer Science UDINUS. his research focus is on data science and natural language processing (NLP).

Sri Winarno, Ph.D
He is a Ph.D in field education sciences and technology. He is senior lecturer at the Infomatics Engineering, Dian Nuswantoro University. His research focuses on education and intelligent system. He can be contacted at email: sri.winarno@dsn.dinus.ac.id.

Adhitya Nugraha, S.Kom, M.CS
He was born in Palangkaraya, Indonesia, in March 1987. He received the Bachelor of Computer Science from Universitas Dian Nuswantoro (UDINUS), Semarang, Indonesia, in 2010, and the Master of Computer Science from Technical Malaysia Melaka University (UTeM),intelligence in 2012. He is currently a Lecturer with UDINUS. His research interests are computer networks and artificial intelligence.

https://doi.org/10.25077/TEKNOSI.v9i3.2023.290-298