Sentiment Analysis of Pro-Israel Product Boycott Action Using IndoBERT Method on Unbalanced Data

Auliana Rizky Puspita Dewi[1], Slamet Riyadi[2*], Cahya Damarjati[3], Nor Ashidi Mat Isa[4], Annisa Divayu Andriyani[5]

[1,2,3]*Department of Information Technology, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia*

[4]*School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Nibong Tebal, Malaysia*

[5]*Faculty of Electrical and Electronics Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Malaysia*

*corr-author: riyadi@umy.ac.id

Abstract - A boycott was an act taken to stop the purchase or use of a particular product or service as a form of public protest against a particular company or group committing a deviation. The Israeli-Palestinian conflict, which had been ongoing since 1948, peaked in October 2023 and had claimed more than 35,000 Palestinian lives. This conflict generated a wide range of public opinions in Indonesia, which were expressed through social media, especially Twitter. Thus, the sentiment analysis of public reactions on Twitter became important to understand the reactions and perspectives of society towards the boycott of Pro-Israel products. This study used the IndoBERT method, which was a variant of the BERT method specifically designed to understand Indonesian. Although many studies had applied the IndoBERT method for sentiment analysis and text classification in Indonesian, none had used the IndoBERT method along with data balancing techniques to analyze Indonesian sentiments regarding the boycott of Pro-Israel products on Twitter. Therefore, this study aimed to develop a sentiment analysis model using the IndoBERT method with more data to examine sentiments related to the boycott of Pro-Israel products on Twitter using imbalanced data, as well as to evaluate the effect of balancing methods using under sampling and oversampling on the model's accuracy and performance. The methods used included data crawling, data preprocessing, labeling with a Lexicon-Based approach, data balancing, and data splitting. The IndoBERT model was trained with 20 epochs, a batch size of 16, and a learning rate of 2e-5. The results of the study showed that the model with balanced data using the oversampling method achieved an accuracy of 97% and an F1-Score of 97%, which was better compared to the model with imbalanced data and the undersampling method. Thus, data balancing using the oversampling method proved to be effective in improving accuracy in sentiment analysis. This research made a significant contribution to understanding the behavior of Indonesian society towards a product boycott supporting Israel and suggested further exploration in parameter optimization and evaluation with larger and more diverse data, as well as further development of data balancing methods to improve the generalization and capabilities of the model.

Keywords: Sentiment analysis; Product boycott; IndoBERT; Balancing; Twitter

I. INTRODUCTION

A boycott is an act taken to stop the purchase or use of a particular product or service as a form of public protest against a particular company or group committing a deviation [1]. This pro-Israeli product boycott may be due to the conflict between Israel and the Palestinians, in which Palestine is under Israeli military occupation [2]. This Israeli-Palestinian conflict has been going on since 1948 and continues to this day [3], where the ceasefire resulted in more than 35,000 Palestinian casualties [4].

The culmination of the Israeli-Palestinian conflict has generated a wide range of public opinions, especially among Indonesians. A variety of sentiments of the Indonesian people are spread through social media platforms, especially on Twitter [5]. Various opinions and reactions were expressed, ranging from sentimental support for product boycott to sentimental opposition to product boycotts [6].

A related study conducted by Tiara Susilawati et al. (2024) on the sentiment of action to boycott Israeli products using the Naïve Bayes classification. This study uses 303 data from twitter about the boycott of Israeli goods, resulting in a 95% accuracy rate with a precision of 96%, recall of 95%, and f1-score of 95%. It can be seen that the method shows a high rate of success in determining the sentiment from the tweets that have been analyzed. However, because the amount of data used is very small, it is possible that the results obtained are less accurate. Therefore, more in-depth training with more data is needed to get better results in more detail and efficiency [7].

Thus, the analysis of public sentiment on social media Twitter becomes essential to understanding the reaction and perspective of the public to the pro-Israeli product boycott action [8]. Text Mining is a process of extracting patterns using information and knowledge from unstructured data. This process focuses on obtaining data patterns, trends, and knowledge hidden in text [9]. Sentiment analysis here is the process used to classify attitudes, opinions, or emotions towards a particular topic that is poured out in the form of text [10]. Using the method of Natural Language Processing (NPL) in sentiment analysis can train computer software to understand text using human language [11]. As for the steps to be taken to the data that is ready to be analyzed, i.e. the pre-processing stage, at this stage, the text data is processed to be ready for analysis by identifying keywords that highlight the core of the text. This processing stage is important for summarizing the text, changing Gaul's sentences, and also for classification [12].

With the advancement of the times, the method of sentimental analysis has evolved quite rapidly, one of which is the IndoBERT method. IndoBERT is a variant of the BERT method specifically designed to understand the Indonesian language. Using a thoroughly trained transformer architecture, IndoBERT can capture text contexts from two directions: left to right and right to left. BERT is trained to model languages and predict next sentences to help models understand language meanings better [13]. To date, many studies have applied the IndoBERT method for sentimental analysis and classification of texts in Indonesian. Therefore, the study examines the performance of a sentiment analysis model using the IndoBERT method in classifying sentiment related to pro-Israel product boycott actions on imbalanced Twitter data. The aim of the study is to develop a model of sentiment analysis using Indo BERT method to analyze sentiment associated with Pro-Israeli product boycotts on unbalanced twitter data, as well as to evaluate the influence of under sampling and oversampling methods on the accuracy and performance of models in balancing data. As well as helping to understand how Indonesian people behave and how they evaluate boycott actions against Pro-Israel products.

Some studies on sentimental analysis have been done before. In a study conducted by Ramanizar et al. (2021) using the Naïve Bayes and Support Vector Machine (SVM) methods, this study obtained accuracy of 74% with the naïve bayes method and 80% for the support vector machine method of 273 data processed, then analyzed using the Lexicon Based approach and identified 67 negative data, 173 neutral data, and 33 positive data. In the study, it used data from Twitter, which is one of the digital platforms that is a means for the public to communicate opinions [14]. Research conducted by Irsyad and Taqwiym (2021) investigated the public's feelings towards the Palestinian people using the Naïve Bayes approach. This study aims to analyze tweet data and classify it with the tools used (Orange). The results processed from such classification showed 56% positive sentiment, 11% negative sentiment, and 33% neutral sentiment with a 76% accuracy rate [15].

Further research conducted by Azhar & Wijayanto (2024) deals with the analysis of public sentiment views on the Palestinian-Israeli conflict and the development of classification techniques using the Support Vector Machine method. (SVM). The stages included data pre-processing, TF-IDF grinding, and the application of the Support Vector Machine algorithm. (SVM). The results showed a high accuracy rate of 88% with a precision of 89%, recall of 88%, and f1-score of 90% [16]. Another study conducted by Makmun, Mustofa, and Turmudi (2022) explains how the Support Vector Machine method is used in analyzing Indonesian citizens' feelings about the Russian-Ukrainian conflict. Datasets obtained from Twitter through the data crawling process using RapidMiner Studio with the keyword “Russian and Ukrainian conflict." With the Naïve Bayes method and feature selection, particle swarm optimization yielded a classification of 55.11% with negative sentiment and 44.89% with positive sentiment [17].

In a study conducted by Mubaraq & Maharani (2022), researchers implemented sentimental analysis of climate change in Indonesia. IndoBERT is a method used in analyzing sentiment, data taken through social media Twitter of 1533 tweets, which then carried out the test process, resulting in an accuracy of 89.6% and a f1-score value of 95.3% using a learning rate of 0.00002 and batch size of 16 [10]. Similarly, Oswari, Yusnitasari, & Wijay (2024) conducted research with the focus being the analysis of the sentimental comments of YouTube users in Indonesia about LGBT. Using IndoBERT fine tuning as its method of classification, accuracy results obtained as much as 75% of the 1493 data obtained through Youtube. Using the Cross-Industry Standard Process for Data Mining model (CRISP-DM) presents a general explanation of the data mining project cycle, with business understanding stages, data understanding, data preprocessing, data labeling, and model implementation [18].

Other research conducted by Imaduddin, Yusfida A'la, and Nugroho (2023) uses the IndoBERT method, which aims to develop a system of sentiment analysis of health app reviews in the Google Play Store. The aim of this system is to help users choose a health app that has the best functionality and service. Using the Python programming language and the Google-Play-Scraper library as a data extraction process, the results were collected from as many as 9310 reviews from a variety of health applications. From the data obtained, 4950 reviews showed positive results and 4360 showed negative results. In the study, 96% accuracy was obtained with 96% re-call and 95% precision [19]. As for the study conducted by Jayadianti et al. (2022), the researchers used the IndoBERT method with Recurrent Convolutional Neural Networks (RCNN) to automate sentiment analysis of Indonesian-language reviews. The data used in this study consisted of positive sentiment, neutral sentiment, and negative sentiment. The results obtained from the test show a high accuracy of 96%; it can be seen that IndoBERT with RCNN can be used in other natural language processing [20].

As discussed in the literature above, there has been a lot of sentiment analysis with a variety of methods and models with a focus on specific data. That's what made me look at public sentiment about the pro-Israeli product boycott that's being talked about on social media, especially Twitter. Choosing such topics to be subject to sentimental analysis using the IndoBERT approach is the right choice to know how the public responds to them.

II. METHOD

This section describes the datasets used and the process of classifying public sentiment against calls for a boycott of products that support Israel. The stages of this research include data crawling on Twitter sites, data preprocessing, data labeling, implementation of IndoBERT methods, and evaluation stages. Fig. 1 is an illustration of the stage of research.

A. Data Crawling

In this study, the data obtained came from the comments of social media users on the Twitter platform by crawling data using the Google Collaboratory. The data collected was based on a set of tweets with the keywords “boycott of Israeli products” and “boycott of pro-Israeli products”. Next, the raw tweet collection is stored in the Comma Separated Values (CSV) format for further processing [21].

Table 1 presents sample tweets extracted during the data crawling process. Each entry includes a timestamp indicating when the tweet was posted and the content of the tweet itself. The dataset primarily consists of user-generated comments expressing opinions related to the boycott of Israeli and pro-Israeli products.

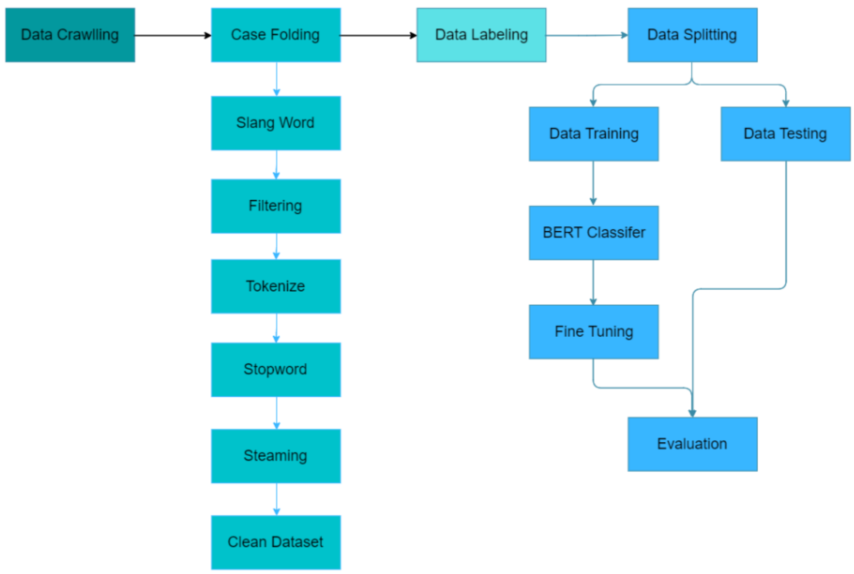

Fig. 1 Research stage

TABLE I
SAMPLE OF THE CRAWLED DATASET

No	Content	
	Waktu	**Ulasan**
1	Wed Feb 21 14:26:57 2024	Tetap Istiqomah boikot seluruh produk terafiliasi zionist.
2	Wed Feb 21 08:15:47 2024	Please terus boikot produk Israel guyssss!!
3	Tue Feb 20 16:38:30 2024	Alhamdulillah masih konsisten boikot produk israel..
4	Thu Nov 16 06:34:05 2023	Boikot produk pro-Israel, jangan cakap saya tidak sensitif – Uqasha https://t.co/tkQipOlluk #Hiburan #aktres
5	Wed Feb 14 12:38:38 2024	btw, tolong tetap boikot brand dan produk pro-israel ya teman-teman.

As seen in TABLE , the dataset contains textual data that may include informal language, abbreviations, and emoticons. Additionally, some tweets contain hyperlinks or hashtags that may require further preprocessing before analysis. The collected dataset will be cleaned and preprocessed in subsequent steps to ensure its suitability for sentiment analysis and other natural language processing (NLP) tasks.

B. Pre-processing Data

Data pre-processing is an important stage in data analysis, involving the preparation and transformation of raw data into a more structured and ready-to-analyze format [22]. As for the phases in data preprocessing to remove special characters and irrelevant symbols, stemming, tokenization, and stopwords [23], here's an explanation of each stage of preprocessing:

1) Case Folding: Converting text with large letters to small letters (lowercase); this is done to facilitate the approach process [24].

2) Slangword: Means a gaul sentence by replacing it with a sentence that corresponds to the correct language principle based on the dictionary.

3) Filtering: Clears the data by removing unnecessary characters [10].

4) Tokenization: Divides the sentence into parts of elements or tokens according to the applicable language norm.

5) Stopword Removal: Removes words that do not add value to the sentiment process [25].

6) Stemming: The process of simplifying words to basic form using stemmer algorithms. For example, "taking" becomes the word "take." It's important in natural language processing to process words broken out of documents [26].

C. Labelling Data

Lexicon-based is a method of labeling in sentiment analysis that uses an approach to determine the sentiment of a text based on the sentiment values of words in the text [18]. In this study, the Inset Lexicon dictionary was used to analyze Indonesian-language data. This Inset Lexicon dictionary contains words that are useful to identify whether a word or sentence has positive polarity, neutral polarity, or negative polarity values. After labeling each word, the sentiment score is calculated by summing up the words with positive and negative values [27]. To calculate sentiment scores, a formula is generally used to sum up the weight of words in a text. Here's the formula used:

$$Sentiment\ Score\ (StSc) = \frac{Positive\ Words - Negative\ Words}{Total\ Words} \quad (1)$$

If a text has a polarity value >0, then the text is considered to be equal to a positive sentiment text; if a text has a polarity value = 0, then that text is deemed to have a neutral sentiment; and if the text has polarity values <0, then it is regarded as having a negative sentiment. Using as many as 7321 pieces of data obtained from crawling through the Twitter platform, that was then analyzed using a lexicon-based approach. As illustrated in Fig. 2, It is known that neutral sentimental sentiments obtained values of 60.25%, or 4411 data, 27.99% for positive sentiments, or 2049 data, and 11.76% for negative sentiments, or 861 data.

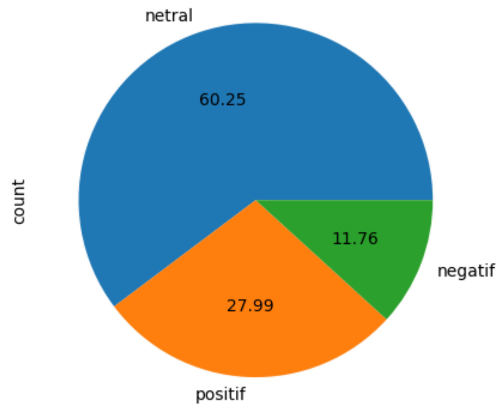

Fig. 2 Labelling results

TABLE presents examples of how sentiment labels were assigned to different tweets. Each row in the table represents a sample tweet along with its corresponding sentiment label. The sentiment classification is determined based on the sentiment score obtained through the lexicon-based approach. For instance, the first tweet expresses support for replacing fast food with local businesses, which is classified as a positive sentiment. The second tweet is a neutral observation without strong positive or negative emotion, thus labeled as neutral. Meanwhile, the third tweet expresses skepticism or criticism toward the effectiveness of boycotts, making it fall under the negative sentiment category. These examples illustrate how the sentiment analysis method assigns labels based on the lexicon-based approach.

D. Balancing Data

In this process, data balancing is carried out using two balancing methods, namely, undersampling and oversampling. The balancing process is useful for overcoming dataset imbalances that aim to improve the ability of classification algorithms.

1) Undersampling Data: The undersampling technique aims to reduce the number of samples from the majority class and equalize with the minority class to balance the original data before the learning process, which ultimately improves classification performance unbalanced data [28].With 861 data for each sentiment data neutral, positive, and negative.

Upsample or Oversample Data: Upsample or oversample is a method used in dataset balancing using the Random Over Sampling (ROS) technique [29]. Random Over Sampling (ROS) involves random duplication of samples in the minority class. This technique utilizes the original data and can support finding the best parameters for the model, with 4411 data for neutral, positive, and negative sentiment data.

E. Spliting Data

The data will be divided into three parts after reaching the pre-processing stage, namely train, validation, and test data. Train data is used to train the model, data validation is used to measure the accuracy of the model in predicting sentiment according to the data to reduce overfitting, and test data is used to evaluate model performance. With data division, 70% for train data, 20% data validation, and 10% for train data.

F. IndoBERT Embedding

The IndoBERT method has the same architecture as BERT, but what distinguishes the two is the dataset used in the pre-training stage. The dataset for training IndoBERT is called Indo4B, which consists of about 4 billion words spread over about 250 million sentences. Indo4B includes sentences in both formal and colloquial Indonesian taken from 15 datasets. Two datasets cover colloquial language, eight cover formal language, and the rest have a mixed style between colloquial and formal language [30]. Bidirectional Encoder Representations from Transformers (BERT) is designed to perform bidirectional training of unlabeled text by co-conditioning each context layer. There are two stages in BERT: pretraining and fine-tuning. In the pretraining stage, the model is trained on unlabeled data with various training tasks. While in the fine-tuning stage, the BERT model is initialized with pre-trained parameters, then all parameters are optimized using labeled data [31].

TABLE II
EXAMPLE SENTENCE LABELLING RESULTS

No	Content	
	Review	**Label**
1	boikot makan ayam goreng mcd kfc dsb ganti menu ayam goreng umkm cerdas	positif
2	wahh sih temenku orang boikot hot post sbux mcd kynya respek sih	netral
3	boikot sih serah tebang pilih ngeboikot saham seuprit banding phk pengurangna karyawan indonesia kayak mcd afiliasi	negatif

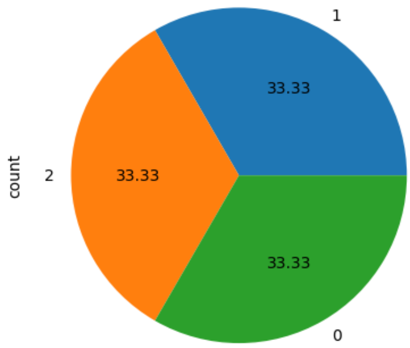

Fig. 3 Balancing result

G. IndoBERT Embedding

The IndoBERT method has the same architecture as BERT, but what distinguishes the two is the dataset used in the pre-training stage. The dataset for training IndoBERT is called Indo4B, which consists of about 4 billion words spread over about 250 million sentences. Indo4B includes sentences in both formal and colloquial Indonesian taken from 15 datasets. Two datasets cover colloquial language, eight cover formal language, and the rest have a mixed style between colloquial and formal language [30]. Bidirectional Encoder Representations from Transformers (BERT) is designed to perform bidirectional training of unlabeled text by co-conditioning each context layer. There are two stages in BERT: pretraining and fine-tuning. In the pretraining stage, the model is trained on unlabeled data with various training tasks. While in the fine-tuning stage, the BERT model is initialized with pre-trained parameters, then all parameters are optimized using labeled data [31].

The initial stage in the embedding process involves adding special tokens at the beginning and end of the sentence, namely CLS at the beginning of the sentence and SEP at the end of the sentence, to distinguish the beginning and end of the sentence. After that, word chunk tokenization is performed, which aims to convert sentences into tokens or words [13]. Fig. 4 illustrates the embedding flow using IndoBERT, which consists of three main embedding components: token embedding, segment embedding, and positional embedding.

In the token embedding stage, each word or subword in the input sentence is converted into a numerical representation based on the pre-trained IndoBERT vocabulary. The figure shows how a sample sentence is tokenized and assigned corresponding token IDs. The segment embedding is useful for distinguishing between different segments of input text, particularly in tasks that involve sentence-pair classification. In the IndoBERT model, if the input consists of only one sentence, then the segment embedding value remains 0 (zero). This process helps the model understand the context and boundaries of the sentence in the text. Lastly, positional embedding is applied to add information about the position of each word in the sentence. Unlike traditional recurrent models, transformers do not process sequences sequentially; instead, positional embeddings provide a sense of word order to the model, allowing it to understand the relative positions of words within a sentence.

Fig. 4 visually represents these embedding processes by showing how an input sentence undergoes tokenization, special token insertion, segment embedding assignment, and positional embedding encoding. This structured representation helps IndoBERT capture syntactic and semantic relationships between words effectively, improving its performance in various downstream NLP tasks [32].

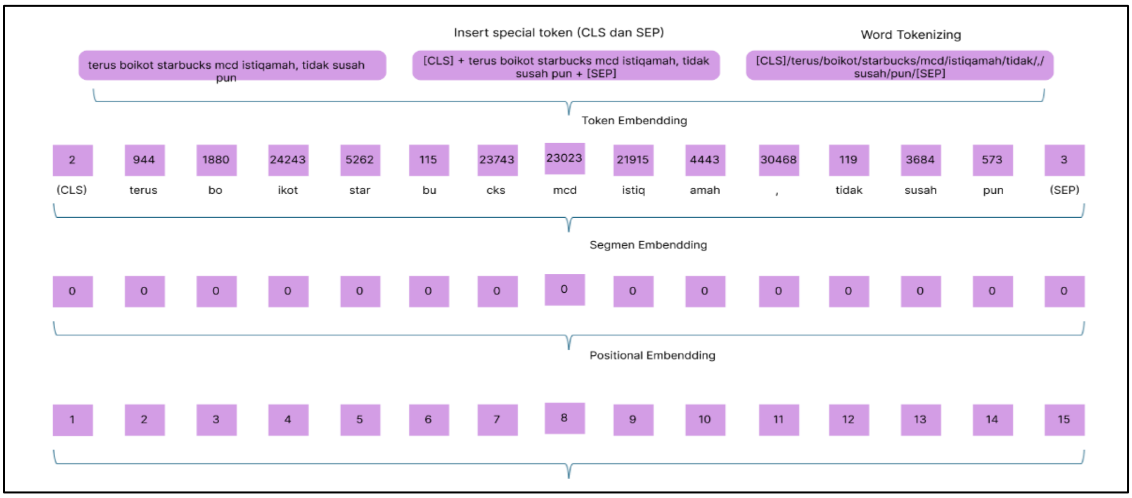

Fig. 4 IndoBERT embedding flow [32]

All experiments were performed with an NVIDIA Tesla T4 GPU with 16 GB of VRAM and 32 GB of system RAM, supplied by the Google Colab Pro environment. The training length for each configuration (20 epochs) differed: roughly 28 minutes for imbalanced data, 22 minutes for undersampling, and 34 minutes for oversampling. Oversampling resulted in a somewhat elevated computing expense due to the augmented data size (13,233 samples), whereas undersampling necessitated the least resources (2,583 samples). Notwithstanding this, all models were trained effectively with a batch size of 16, and no memory complications were experienced.

H. Evaluation

Evaluation is a technique used to measure the classification ability of the model. This research uses a confusion matrix for model evaluation, which produces four key values: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Several metrics such as accuracy, precision, recall (sensitivity), specificity, and F1-score are used to assess the applied model [33].

The accuracy metric measures the ability of the classification model to correctly predict each case in the dataset [34]. It is calculated by comparing the number of correct predictions (both positive and negative) to the total number of predictions made by the model. The accuracy formula is defined as follows:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{2}$$

While accuracy provides an overall measure of model performance, it is important to consider other metrics such as Precision, Recall (Sensitivity), and F1-Score, especially when dealing with an imbalanced dataset. The formulas for these metrics are as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{3}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{4}$$

$$\mathrm{F1-Score} = \frac{2\ x\ (Precision\ x\ Recall)}{Recall+Precision} \tag{5}$$

To better illustrate the concept of TP, TN, FP, and FN in this study, Table III presents the confusion matrix used for evaluating the model’s sentiment classification performance.

In this study, True Positive (TP) refers to cases where the model correctly predicts a positive sentiment, while True Negative (TN) represents instances where the model correctly identifies a negative sentiment. False Positive (FP) occurs when the model incorrectly classifies a negative sentiment as positive, and False Negative (FN) happens when the model fails to detect a positive sentiment and classifies it as negative instead.

By analyzing these values, the evaluation metrics provide insights into how well the model performs in correctly classifying sentiments within the dataset. A high accuracy alone is not always sufficient, especially when dealing with imbalanced data, which is why precision, recall, and F1-score are also crucial in assessing the classification model's robustness and reliability [35].

III. RESULT AND DISCUSSION

This study utilizes the pre-trained IndoBERT method. Indonesia Bidirectional Encoder Representations from Transformers (IndoBERT) was developed using the PyTorch Framework, with transformers based on Bert Base with 12 hidden layers, specifically designed to handle tasks in Indonesian [19]. This study uses 20 epochs for the training process, with a batch size of 16, and with a learning rate of 2e-5 for Adam optimization. Based on Fig. 5, Fig. 7, and Fig. 9, the validation accuracy trend curve continues to rise to the right, and the training loss and validation loss continue to decline.

TABLE III
CONFUSION MATRIX FOR SENTIMENT CLASSIFICATION

Actual/Predicted	**Positive/Negative**	
	Positive (Predicted)	**Negative (Predicted)**
Positive (Actual)	True Positive (TP)	False Negative (FN)
Negative (Actual)	False Positive (FP)	True Negative (TN)

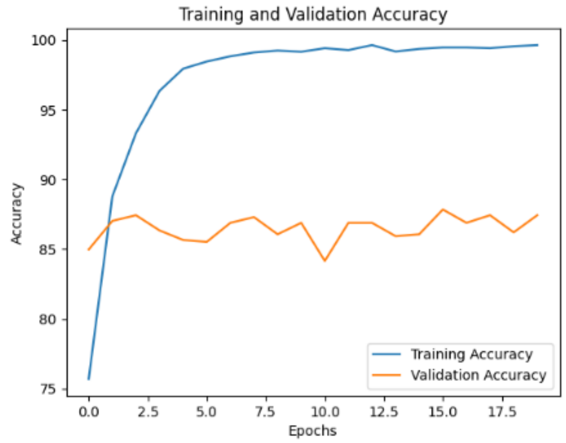

Fig. 5 Training accuracy imbalance

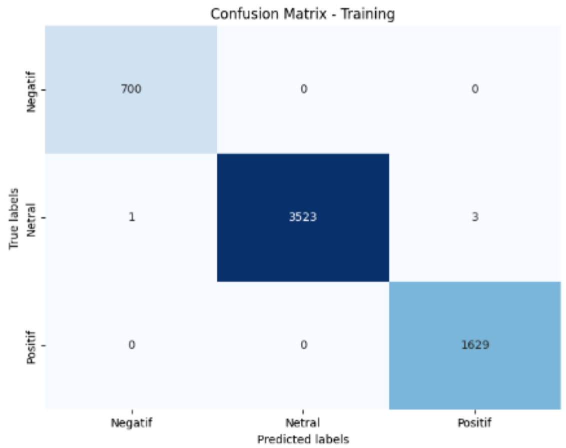

Fig. 6 Confusion matrix imbalance

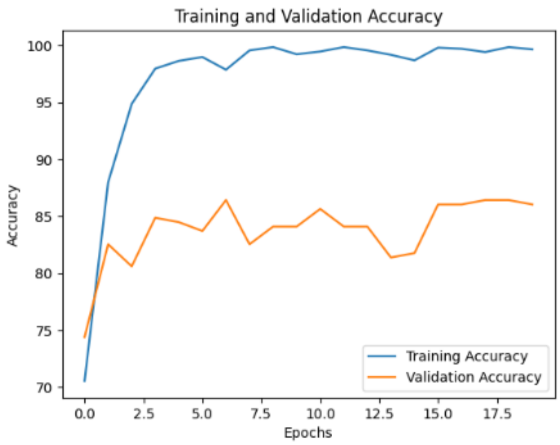

Fig. 7 Training accuracy undersampling

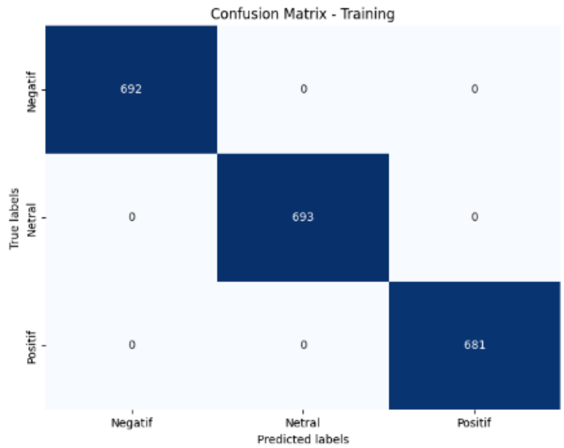

Fig. 8 Confusion matrix undersampling

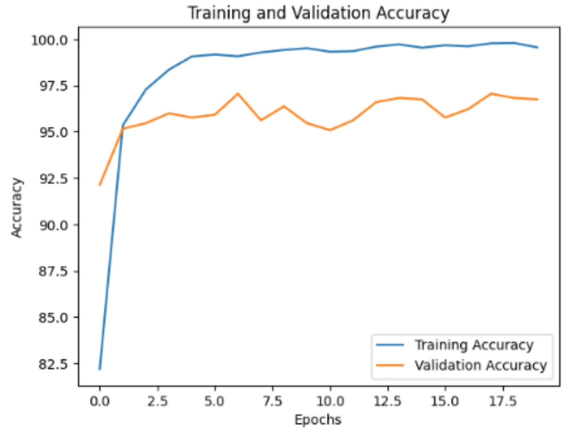

Fig. 9 Training accuracy oversampling

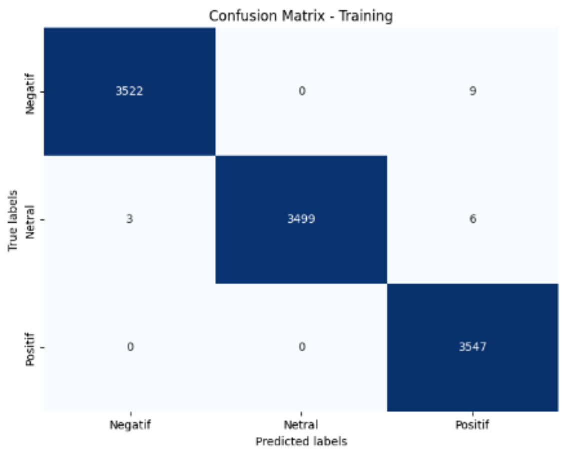

Fig. 10 Confusion matrix oversampling

Fig. 5, Fig. 7, and Fig. 9 respectively show the training and validation accuracy curves for models trained on imbalanced, undersampled, and oversampled datasets. In Fig. 5, the model trained on imbalanced data exhibits a significant gap between training and validation accuracy, indicating overfitting. Similarly, Fig. 7 (undersampling) also shows misalignment between training and validation curves, although with slightly improved performance. In contrast, Fig. 9 (oversampling) presents the most stable and aligned accuracy trends, suggesting better generalization and less overfitting. This highlights that oversampling significantly enhances the model's performance.

Further evaluation using the confusion matrix is presented in Fig. 6, Fig. 8, and Fig. 10. Fig. 6 (imbalanced) shows that the model predicted 1629 positive, 3523 neutral, and 700 negative tweets correctly, with one misclassification from neutral to negative. Fig. 8 (undersampling) demonstrates that the model achieved perfect classification for each sentiment class (692 negative, 693 neutral, and 681 positive), suggesting high performance on the balanced but reduced dataset. Fig. 10 (oversampling) shows strong classification performance as well, with 3547 correctly predicted positive tweets, 3499 neutral, and 3522 negative. Minor misclassifications include 9 negative tweets predicted as positive, 6 neutral as positive, and 3 neutral as negative.

The learning rate of 2e-5 and a batch size of 16 employed in this work were chosen according to established protocols in IndoBERT fine-tuning and corroborated through preliminary tests. Multiple learning rates (5e-5, 3e-5, and 2e-5) were evaluated, with 2e-5 yielding the most consistent training and optimal validation outcomes across experiments, particularly with oversampled data. These parameters correspond with those often employed in prior IndoBERT implementations and were appropriate for our computational resources and dataset dimensions.

In TABLE , the model with imbalance data can produce an accurate value of 0.87 with an F1-Score 0.84 where these results are still unsatisfactory. Then balancing the data using the under-sampling method, the accuracy becomes 0.85 with F1-Score 0.84, but the result is still unsatisfactory. Because the results are still unsatisfactory, data balancing is then carried out again using a different method, namely the oversampling method which produces an accuracy value of 0.97 with an F1-Score value of 0.97, where the treatment succeeds in increasing the accuracy value and F1-Score value.

TABLE IV
PERFORMANCE EVALUATION RESULTS

Model	**Result**		
	Approach	**Accuracy**	**F1-Score**
IndoBERT	Imbalance	0.87	0.84
	Undersamling	0.85	0.84
	Oversampling	0.97	0.97

In contrast to earlier research employing Naïve Bayes and SVM [14], which indicated accuracy rates ranging from 74% to 95%, and other IndoBERT-based sentiment analyses that attained accuracies between 75% and 89.6% [18], our optimal model utilizing IndoBERT and oversampling achieved an accuracy and F1-score of 97%, signifying a notable enhancement due to proficient balancing and fine-tuning methodologies.

IV. CONCLUSION

This research uses transfer learning technique with IndoBERT model that has been trained to analyze sentiment towards Indonesian text. The dataset used for this research consists of 7321 reviews, which are labeled positive, neutral, and negative, respectively. During the training process, 20 epochs and Adam optimization were used, with a learning rate of 2e-5. From the data that has been obtained, a comparison has been made using three treatments, namely IndoBERT ImBalance, IndoBERT Undersampling, and IndoBERT Oversampling. From the results of the treatment that has been carried out, IndoBERT with Oversampling method data produces a better value than the other two methods, with accuracy reaching 0.97, and with an average F1-Score of 0.97. This shows that data balancing treatment using the oversampling method is a good way to improve accuracy in sentiment analysis. To improve the performance of the model, future research can explore parameter optimization, conduct evaluations using larger and more diverse data, and can further develop data balancing methods to improve the generalizability and capabilities of the model.

ACKNOWLEDGEMENT

This experience has truly been one of growth and learning. Universitas Muhammadiyah Yogyakarta supported my endeavor from inception to completion, opening doors to resources and providing a place conducive to focus. For this, I am deeply thankful. Too, I appreciate the supervising lecturers who so generously gave guidance and perspective, helping steer my efforts to fruition. This achievement stands as a testament to their dedication in sharing expertise that expanded my understanding greatly. It brings me joy to reflect on this journey and all who accompanied me along the way. This conclusion emerges from their investment and my good fortune in counting them as mentors. May this research contribute significantly to the advancement of knowledge and serve as a tangible reflection of Universitas Muhammadiyah Yogyakarta's dedication to community service.

REFERENCES

[1] B. Palacios-Florencio, M. Á. Revilla-Camacho, D. Garzón, and C. Prado-Román, "Explaining Boycott Behavior: Proposal and validation of a conceptual model," *Journal of Consumer Behaviour*, vol. 20, no. 5, pp. 1313–1325, Sep. 2021, doi: 10.1002/cb.1937.

[2] D. Suryani Sentosa and N. I. Sitepu, "Fatwas and the Urgency of Compliance," 2024.

[3] T. Farhat, S. Ibrahim, Z. Abdul-Sater, and G. Abu-Sittah, "Responding to the Humanitarian Crisis in Gaza: Damned if You do. Damned if You don't.," *Ann Glob Health*, vol. 89, 2023, doi: 10.5334/aogh.3975.

[4] AJLabs, "Israel-Gaza war in maps and charts: Live tracker," May 2024.

[5] A. Tiara Susilawati, N. Anjeni Lestari, and P. Alpria Nina, "Analysis of Public Sentiment on Twitter Toward Boycott of Israeli Products Using Naïve Bayes Method," *Jurnal Ilmiah Mahasiswa*, vol. 2, no. 1, pp. 26–35, 2024, doi: 10.59603/niantanasikka.v2i1.240.

[6] Y. Matalon, O. Magdaci, A. Almozlino, and D. Yamin, "Using Sentiment Analysis to Predict Opinion Reversals in Political Communication Tweets," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/s41598-021-86510-w.

[7] M. Hamka and D. Ratna Sari, "Sentiment Analysis And Information Extraction Of Online Learning Using Lexicon Approach," 2022.

[8] S. Damayanti, L. Dwi Oktaviana, and T. Bratakusuma, "Data Mining Using Support Vector Machine Model for Baturraden Tourism Visitor Satisfaction Prediction," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, 2024, doi: 10.30865/mib.v8i2.7346.

[9] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text Mining in Big Data Analytics," *Big Data and Cognitive Computing*, vol. 4, pp. 1–34, Mar. 2020, doi: 10.3390/bdcc4010001.

[10] M. F. Mubaraq and W. Maharani, "Sentiment Analysis on Twitter Social Media towards Climate Change on Indonesia Using IndoBERT Model," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, p. 2426, Oct. 2022, doi: 10.30865/mib.v6i4.4570.

[11] R. G. Katryn, "Text Preprocessing: Tahap Awal dalam Natural Language Processing (NLP)," Mandiri Engineering. Accessed: May 17, 2024. [Online]. Available: https://medium.com/mandiri-engineering/text-preprocessing-tahap-awal-dalam-natural-language-processing-nlp-bc5fbb6606a

[12] Amazon, "What is Sentiment Analysis?," Amazon. Accessed: May 23, 2024. [Online]. Available: https://aws.amazon.com/id/what-is/sentiment-analysis/

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1810.04805

[14] H. Ramanizar, A. Fajri, R. Binsar Sinaga, H. Mubarok, A. D. Pangestu, and D. S. Prasvita, *Sentiment Analysis of Twitter Users on the Conflict between Palestine and Israel Using Naïve Bayes Classification and Support Vector Machine Methods*. 2021.

[15] H. Irsyad and A. Taqwiym, "Analyzing People's Sentiment towards Palestinians with Naive Bayes Classification," 2021.

[16] R. Azhar and M. F. Wijayanto, "Sentiment Analysis on Twitter: Uncovering Perceptions and Public Emotions Around the Palestinian-Israeli Conflict," 2024.

[17] M. Makmun, Z. Mustofa, and A. Turmudi, "Bulletin of Information Technology (BIT) Sentiment Analysis of Indonesian People in Russia-Ukraine Conflict on Twitter," vol. 3, no. 4, pp. 355–366, 2022, doi: 10.47065/bit.v3i1.

[18] T. Oswari, T. Yusnitasari, and S. Wijay, "Sentiment Analysis of Indonesian YouTube Reviews About Lesbian, Gay, Bisexual, and Transgender (LGBT) using IndoBERT Fine Tuning," vol. 15, no. 1, pp. 2088–1541, 2024, doi: 10.24843/LKJITI.2024.v15.i01.p03.

[19] H. Imaduddin, F. Yusfida A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," 2023. [Online]. Available: www.ijacsa.thesai.org

[20] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "Sentiment analysis of Indonesian reviews using IndoBERT and R-CNN refinement," *ILKOM Jurnal Ilmiah*, vol. 14, pp. 348–354, Dec. 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

[21] N. Mahfudiyah and A. Alamsyah, "Consumer Perception Analysis Of Gojek Service Quality Using Sentiment Analysis And Topic Modeling Based On Deep Learning IndoBERT," 2022.

[22] R. Merdiansah and A. Ali Ridha, "Sentiment Analysis of Indonesian X Users Regarding Electric Vehicles Using IndoBERT," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI*, vol. 7, no. 1, pp. 221–228, 2024.

[23] Deepanshi, "Text Preprocessing in NLP with Python Codes," Analytics Vidhya.

[24] H. M. Lee and Y. Sibaroni, "Comparison of IndoBERTweet and Support Vector Machine in Sentiment Analysis of Racing Circuit Development in Indonesia," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 1, pp. 99–106, 2023, doi: 10.30865/mib.v7i1.5380.

[25] M. Ali, "NLTK Sentiment Analysis Tutorial for Beginners," datacamp. Accessed: May 23, 2024. [Online]. Available:

https://www.datacamp.com/tutorial/text-analytics-beginners-nltk

[26] J. Murel and E. Kavlakoglu, "What is stemming?," IBM. Accessed: May 23, 2024. [Online]. Available: https://www.ibm.com/topics/stemming

[27] Y. Fauziah, B. Yuwono, and A. S. Aribowo, "Lexicon Based Sentiment Analysis in Indonesia Languages : A Systematic Literature Review," *RSF Conference Series: Engineering and Technology*, vol. 1, no. 1, pp. 363–367, Dec. 2021, doi: 10.31098/cset.v1i1.397.

[28] M. Zheng *et al.*, "UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification," *Inf Sci (N Y)*, vol. 576, pp. 658–680, Oct. 2021, doi: 10.1016/j.ins.2021.07.053.

[29] S. Fulvi Intan *et al.*, "Perbandingan Algoritma KNN, NBC, dan SVM: Analisis Sentimen Masyarakat Terhadap Perparkiran di Kota Pekanbaru," 2023.

[30] B. Wilie *et al.*, "IndoNLU: Benchmarks and Resources for Evaluating Indonesian Natural Language Comprehension," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[31] S. Dharmawan,) Viny, C. Mawardi,) Novario, and J. Perdana, "Jurnal Ilmu Komputer dan Sistem Informasi Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT)," 2023.

[32] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 1071–1078, Feb. 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[33] S. Riyadi, A. Divayu Andriyani, and S. Noraini Sulaiman, "Improving Hate Speech Detection Using Double-Layers Hybrid CNN-RNN Model on Imbalanced Dataset," *IEEE Access*, vol. 12, pp. 159660–159668, 2024, doi: 10.1109/ACCESS.2024.3487433.

[34] S. Riyadi, F. Alfiana, M. I. Al-Habib, C. Damarjati, and E. Loniza, "Early Detection of Heart Disease on Imbalanced Data Using Gradient Boosting," *2024 Ninth International Conference on Informatics and Computing (ICIC)*, pp. 1–6, Oct. 2024, doi: 10.1109/ICIC64337.2024.10956329.

[35] S. Riyadi, A. D. Andriyani, A. M. Masyhur, C. Damarjati, and M. I. Solihin, "Detection of Indonesian Hate Speech on Twitter Using Hybrid CNN-RNN," *Proceeding - International Conference on Information Technology and Computing 2023, ICITCOM 2023*, pp. 352–356, 2023, doi: 10.1109/ICITCOM60176.2023.10442041.