ISSN 2085-4552

Sentiment Analysis in E-Commerce: Beauty Product Reviews

Gavrila Louise Tumanggor[1], Feliks Victor Parningotan Samosir[2]
[1, 2] Informatics Study Program, Faculty of Computer Science and Information Technology, Indonesia
[1]01082210011@student.uph.edu [2]feliks.parningotan@uph.edu

Accepted 19 August 2024
Approved 08 January 2025

***Abstract*— The increasing popularity of online shopping platforms is fueling the need for automated sentiment analysis for product reviews. This research aims to build an automatic sentiment analysis model in Indonesian for e-commerce product reviews. This model is expected to help consumers make purchasing decisions more quickly. We utilize the IndoBERT model, which has shown to be quite effective for general sentiment analysis, achieving an evaluation accuracy of 66.2% despite a high evaluation loss of 0.8006. The approach used combines Natural Language Processing (NLP) and Machine Learning (ML) techniques. It is hoped that this research will be useful for consumers, shop owners, and researchers in efficiently understanding the sentiment of e-commerce product reviews.**

***Index Terms*— *E-Commerce, Sentiment Analysis, Product Reviews, IndoBERT*.**

I. INTRODUCTION

In this digital era, e-commerce has become an increasingly popular platform for online shopping, with marketplaces like Shopee in Indonesia offering users the convenience of conducting transactions online. Product reviews on e-commerce platforms such as Shopee serve as crucial sources of information for consumers to assist them in making purchasing decisions[1]. These reviews encompass a range of opinions, from positive to negative, and neutral sentiments, providing insights into the quality of a product and significantly impacting both consumers and producers. In today's business environment, customer satisfaction stands as a primary goal for companies, and customer reviews posted on social media can greatly influence the perceptions of potential customers. The increasing prevalence of sentiments and opinions on product review sites underscores the importance of having review data available for analyzing customer opinions [2].

Natural Language Processing (NLP) is a computer science field that focuses on the interaction between computers and human language. NLP techniques can be used to process and analyze text, such as tokenization, stemming, lemmatization, and text classification. Sentiment analysis is a technique used to identify and understand opinions or feelings contained within text [2]. In the context of e- commerce, sentiment analysis of product reviews can help various parties, such as helping consumers choose the right product by understanding other users' opinions and experiences with the product [3], assist store owners in understanding how their products are received by consumers, identify product deficiencies, and improve the quality of their products and services [4], also helps researchers in understanding trends and patterns in consumer opinions towards certain products [5].

Manual sentiment analysis of e-commerce product reviews takes a lot of time and effort, especially if the number of reviews is very large[6]. This is a challenge, especially for consumers who want to quickly find out the general sentiment towards a product. Therefore, an automatic sentiment analysis technique is needed that can process and analyze product reviews quickly and accurately.

Machine learning (ML) is a sector of artificial intelligence dedicated to employing algorithms systematically to uncover the inherent connections within data and information [7]. For instance, ML systems can undergo training with automatic speech recognition systems (like iPhone's Siri) to transform acoustic data into a sequence of speech into a semantic framework represented by a string of words ML algorithms can be used to build sentiment analysis models that can predict the sentiment of a text.

This research aims to develop an automatic sentiment analysis model for e-commerce product reviews in Indonesian. This model is expected to help consumers make purchasing decisions more quickly and easily.

Product reviews play an important role for shop owners and buyers. For store owners, reviews provide direct feedback on the quality and performance of their products. By understanding what customers like or dislike, they can improve their products to meet market needs. Positive reviews also build customer trust and a store's reputation, while negative reviews provide an opportunity to respond quickly to any issues that may arise. For buyers, reviews provide valuable

guidance for making better purchasing decisions. They can learn about other people's experiences with the product before they decide to buy it. Reviews also provide validation of product claims and help buyers avoid waste by uncovering any problems or shortcomings they may encounter[8].

II. THEORY

There are several studies that have carried out e-commerce sentiment analysis. First, in the paper entitled "Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT" by Chandradev, et al [9]. The research in this paper aims to analyze the sentiment of hotel reviews using the BERT deep learning method. The model used in this research is SmallBERT, which is a variant of BERT with a smaller size but still maintaining language processing capabilities. The use of the SmallBERT model in this research involved a fine-tuning process on a dataset of 515k hotel reviews. The sentiment analysis process begins by loading the model resulting from the fine-tuning that has been carried out previously. Next, the tokenization stage is carried out using the BERT Tokenizer to convert the review text into a number representation that matches the BERT vocabulary. After that, input packing is carried out to change the data structure into input that is in accordance with the BERT model. The model then performs sentiment classification with binary classification into negative and positive labels. The model achieves low loss values and high accuracy on training and validation data. The model does not suffer from overfitting or underfitting, indicating good ability in analyzing hotel review sentiment. The advantage of this research is the efficient use of the SmallBERT model in analyzing hotel review sentiment with high accuracy. In addition, the BERT deep learning approach makes it possible to capture relationships between words contextually, improving the model's performance on complex sequential tasks in natural language processing A shortcoming of this study may lie in the scale of the dataset used. The results of this research include achieving loss values and accuracy of the SmallBERT model on training and validation data, as well as visualization of sentiment comparison data using bar charts and pie charts.

Further research is in the paper entitled "Implementasi Fine-Tuning BERT untuk Analisis Sentimen terhadap Review Aplikasi PUBG Mobile di Google Play Store" by Alex Sander Prasetya Braja, and Achmad Kodar[10]. In the research discussed in this paper, sentiment analysis was carried out on user reviews of the PUBG Mobile application on the Google Play Store. This research uses the BERT BASE Multilingual and IndoBERT BASE models. Both models are fine-tuned with predetermined hyperparameters. These models were trained using PyTorch and the transformers package from hugging face. The research results show that IndoBERT BASE obtained the best testing accuracy on batch size 32 of 0.93519, while BERT BASE Multilingual obtained the best testing accuracy at 0.71149. Using a fine-tuning approach allows the model to learn from larger data in Indonesian, including formal language and slang. However, the drawback may lie in the unbalanced distribution of sentiment classes in TextBlob-based labeling data.

The next research is regarding "Implementasi-Bert-Untuk-Analisis-Sentimen-Terhadap-Ulasan-Aplikasi-Flip-Berbahasa-Indonesia" [11]. This research aims to apply the Language-Based Transformer (BERT) model in analyzing the sentiment of user reviews of the Flip application in Indonesia. Through natural language processing (NLP) and tokenization techniques, user review data is analyzed using the BERT model to determine positive, negative or neutral sentiment. The research results show that the BERT model can be effective in analyzing sentiment, with the majority of reviews being positive, temporary Negative reviews provide input for application improvement. In this research, it can be seen that BERT implementation has the potential to increase understanding of user sentiment towards the Flip application and provide valuable insights for developers in improving service quality.

The next paper entitled "Analisis Sentimen Pada Ulasan Aplikasi Tokopedia Menggunakan Klasifikasi Naïve Bayes" by Aurelia, et al[12]. The research discussed in this paper conducts sentiment analysis on user reviews of the Tokopedia application using the Naïve Bayes classification method. Data were collected from the Google Play Store and labeled as positive or negative based on user-provided scores. Subsequent processes involved data preprocessing, including the creation of a corpus from reviews and the formation of a sparse document-term matrix. The data were then split into training and testing sets for the classification process. The Naïve Bayes method was chosen as the classification model for its effectiveness in sentiment analysis, particularly in the context of text mining. This model can classify reviews as positive or negative based on the words they contain. The research achieved an accuracy rate of 82.97%, demonstrating success in classifying reviews. The advantage of Naïve Bayes lies in its ability to achieve high accuracy with a relatively small amount of training data. However, its drawback may lie in the naive assumption of feature independence, which may not always hold true in real-world data. Out of 1819 analyzed reviews, 338 were classified as positive and 1481 as negative. For future research, it is suggested to conduct a deeper analysis of neutral reviews (with a score of 3) that were excluded in this study. Additionally, other classification methods could be considered to compare their effectiveness with Naïve Bayes. Thus, this research successfully implements the Naïve Bayes model for sentiment analysis of Tokopedia application reviews with a satisfactory level of accuracy.

The fifth paper is entitled “Analisis Sentimen AicoGPT (Generative Pre-trained Transformer) Menggunakan TF-IDF” by Rahayu, et al [13]. The aim of the study presented in this paper is to conduct sentiment analysis on the reviews of the AicoGPT application from the Google Play Store. The research method involves data collection of AicoGPT comments, data preprocessing, modeling using classification with Logistic Regression (LR) and Support Vector Machine (SVM) utilizing the TF-IDF technique, and evaluation of results using confusion matrix and ROC Curve. The models utilized in this research are Logistic Regression (LR) and Support Vector Machine (SVM) employing the TF-IDF technique. LR is utilized for data classification, while SVM is used to compare model performance. Evaluation results using ROC Curve indicate that the SVM model performs the best with an f1-score of 73.34%. The strength of this study lies in the utilization of advanced methods such as TF-IDF and classification models which can provide valuable insights for AicoGPT application developers. However, its weakness may lie in the limitation of the dataset used and the classification model that does not encompass various possible sentiments that may arise in reviews. The findings of this research indicate that the SVM model outperforms LR in conducting sentiment analysis on AicoGPT application reviews. Thus, this study successfully achieves its goal of providing input for application developers and users. For future research, a more in-depth analysis of specific sentiment types, the utilization of more complex classification models, and the use of broader and more representative datasets to enhance the accuracy and generalization of sentiment analysis results could be considered.

The last paper entitled “Analisis Sentimen pada Review Pengguna E-Commerce Menggunakan Algoritma Naïve Bayes” by Abdul, et al[14]. In this study, sentiment analysis was conducted on product reviews from Kinophonecell store on the Shopee e-commerce platform using the Naïve Bayes classification algorithm. The Naïve Bayes model was employed to predict the sentiment category of customer reviews towards the products, classifying the reviews as either positive or negative based on the available dataset. The results of the study indicate that the Naïve Bayes method yielded effective outcomes, with an accuracy of 99.5%, precision of 99.49%, and recall of 100%. Although the accuracy did not reach 100%, these results are considered satisfactory for predicting product review sentiments. One advantage of this research is the utilization of the Naïve Bayes method, which is relatively easy to implement and provides good results in predicting product review sentiments. However, a limitation is the accuracy not reaching 100%, suggesting the need for improving the quality of the training data to enhance prediction accuracy. The study also revealed that positive sentiments outweighed negative sentiments, with a significantly larger number of positive sentiments compared to negative ones. The quantity of training data significantly influences the system's predictions, underscoring the importance of high-quality training data. For future research, it is recommended to enhance the quality of training data by increasing the dataset size and refining ambiguous sentiment labels. Additionally, employing alternative classification methods for comparison with Naïve Bayes results could be beneficial. Conducting deeper analyses on factors influencing product review sentiments and integrating sentiment analysis with other factors such as pricing, promotions, or service quality would provide a more comprehensive understanding.

Based on the literature review, it was decided to use BERT due to its advantages in understanding the context and relationships between words in text, which is crucial for sentiment analysis. Previous studies, such as the research by Chandradev et al. that used SmallBERT for sentiment analysis of hotel reviews and demonstrated high accuracy without overfitting or underfitting, have proven the effectiveness of this model. Additionally, the research by Alex Sander Prasetya Braja and Achmad Kodar, which employed IndoBERT BASE for sentiment analysis of PUBG Mobile application reviews, also showed impressive results with a testing accuracy of 0.93519. Specifically, we will use IndoBERT because our dataset is in Indonesian, and this model has been optimized to understand the nuances and variations of the Indonesian language, including the use of formal and slang expressions. The use of IndoBERT is expected to improve the accuracy of sentiment predictions and provide deeper insights into e- commerce product reviews in the context of the Indonesian language.

III. METHOD

A. Data Preparation

First, data was collected from Hugging Face Hub, a platform that provides open access to a variety of datasets and NLP models that can be used for a variety of tasks, including sentiment analysis. In this research, the review dataset used is from Sekar Mulyani [15] namely beauty products from Shopee taken from the Hugging Face Hub using the API or method provided by Hugging Face.

	Review	Bintang 1	Bintang 2	Bintang 3	Bintang 4	Bintang 5
0	terima kasih shoppe paket ny udah datang denga...	True	False	False	False	False
1	kecewa sekali box nya kurang rapi dan produkny...	False	True	False	False	False
2	it's been days and the itch still haven't gone...	False	True	False	False	False
3	bukanny makin mulus malah tumbuh jerawat gede ...	False	True	False	False	False
4	pengalaman penggunaan:gampang patah	False	True	False	False	False
...	...	...	...	...	...	...
57187	free beauty blender tidak ada. tadi nya saya p...	False	False	False	True	False
57188	kalian wajib liat vid ini.	False	False	True	False	False
57189	pesanan tidak sesuai, tidak tepat waktu, tidak...	False	True	False	False	False
57190	terimakasih, barang sdh di terima dengan baik...	False	False	False	True	False
57191	isi tidak sesuai	True	False	False	False	False

57192 rows × 6 columns

Fig. 1. Initial Dataframe

Based on figure 1, the data used has 57.192 rows and 6 columns, namely review column, 1 star column, 2star column and so on up to 5star column. In this column, if an item gets 1 star, then the 1star column will be true while the other stars will be false.

	Review	Bintang 1	Bintang 2	Bintang 3	Bintang 4	Bintang 5
0	semua bagus, nyampe langsung coba tapi koq bau...	False	False	True	False	False
1	gampang dipake buat ngebentuk mata bahkan bagi...	False	False	False	False	True
2	baru beli udah kering, gada video unboxing di ...	True	False	False	False	False
3	efektivitas:oke tekstur:oke kandungan:oke pesa...	True	False	False	False	False
4	pengiriman ok, hny saja barang tidak di tambah...	False	False	False	True	False
...	...	...	...	...	...	...
2495	parfum nya nyengat dan di kulitku kering bgt, ...	False	True	False	False	False
2496	produk okeh tapi pengirimannya lama banget, ka...	False	True	False	False	False
2497	saya pesen paket yang awet muda, harusnya dape...	True	False	False	False	False
2498	sdh di terima tp belum sempat di coba	False	False	False	True	False
2499	gak bisa pakainya	False	False	False	True	False

2500 rows × 6 columns

Fig. 2. Dataframe used

Figure 2 shows 2500 data points were randomly selected from the original dataset due to limitations in computational resources, such as processing power, memory, and time constraints. This random selection ensures that the subset remains unbiased and representative of the overall data distribution, capturing the essential characteristics and diversity of the original dataset. By working with a smaller, randomly chosen subset, the model can be trained more efficiently, allowing for faster iterations and testing of different parameters, while still maintaining the integrity of the data within the given resource limitations.

Then identify and handle missing values. This activity is carried out to maintain data quality. If this is not done, missing values can result in deviations or distortions that may occur in the analysis results due to certain characteristics of the data used and errors in the analysis, as well as reducing the accuracy of the results. By identifying and handling missing values, the data becomes more complete and can provide more reliable analysis results. The missing values contained in the data used are all 0 so there is no need to clean them.

Sentiment	count
Negative	1066
Positif	886
Neutral	548

dtype: int64

Fig. 3 Data Labeling

Entering data labeling as you can see at Figure 3, sentiment or ratings from product reviews are labeled based on a star scale given by users (for example: 1 star for negative sentiment, 3 stars for neutral sentiment, and 5 stars for positive sentiment). These labels are then used as a basis for training the sentiment analysis model.

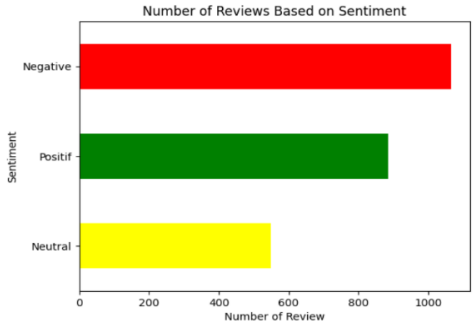

Fig. 3 Results reviews from the new sentiment

has a total of 1066 reviews, positive (4 and 5 stars) which has 886, and neutral which has 548 reviews. Neutral sentiment has the least number of reviews because only 3 stars. We can see the graphics of resuls reviews from the new sentiment in figure 4.

Each product review is divided into tokens or small units (usually words or sub-words) using a tokenizer, in this case, the tokenizer used is nltk Tokenizer. The tokenization process allows the text representation to be input that can be understood by the NLP model.

Entering the data cleaning and data preparation stage, at the data cleaning and preparation stage, text cleaning is carried out which includes removing stop words and normalizing the text to improve the quality of the analysis. The stopwords used are a collection of common words in Indonesian that tend not to have important meaning in the context of text analysis, such as "and", "or", and the like. After removing stopwords, empty tokens are cleaned to ensure data cleanliness. However, there is a problem when some words that are actually irrelevant are still read. This is caused by the existence of abbreviations that are not affected by the stopword remover.

```
30 Kata dengan Frekuensi Tertinggi setelah Pembersihan Kedua:
[('cocok', 1653), ('bagus', 1317), ('barang', 1263), ('beli', 1145), ('tekstur',
1054), ('produk', 1049), ('pengiriman', 1000), ('warna', 749), ('pas', 700),
('kecewa', 698), ('aman', 653), ('dikirim', 647), ('performa', 627), ('kulit',
622), ('kemasan', 598), ('pake', 566), ('cepat', 563), ('coba', 515),
('packing', 509), ('warnanya', 443), ('loudlycryingface', 415), ('pakai', 404),
('harga', 390), ('pesanan', 383), ('ok', 381), ('udh', 380), ('bibir', 378),
('kasih', 352), ('sampe', 351), ('kering', 350)]
```

Fig. 4 Second Cleaning Stage

Figure 5 shows the second cleaning stage. Therefore, the second cleaning stage is carried out by excluding certain predetermined words, such as 'nya', 'yang', 'ga', and others contained in the `exclude_words` list. This step aims to increase the accuracy of the analysis by eliminating words that are irrelevant and do not have significant meaning in the context of text analysis.

After data preparation, visualization is used to understand the distribution of sentiment labels in the dataset. Graphs or plots can be used to display the number of reviews for each sentiment category (negative, neutral, positive), helping in visually understanding the distribution of the data. This data

preparation aims to prepare a product review dataset for sentiment analysis, including steps such as collection, cleaning, normalization, and proper data representation for effective sentiment analysis model training.

B. Model Design

Once the data is ready, the IndoBERT model is configured to be used for training. This model is supplied with a tokenizer that has been pre-trained to optimally process text in Indonesian. The detailed model used in the code above is IndoBERT, which is a variant of the BERT model that has been pretrained for the Indonesian language. IndoBERT uses a Transformer architecture like BERT in general, but has been trained specifically for the Indonesian language. This model has parameters that have been pretrained on common natural language understanding (NLP) tasks such as mapping words into vector representations, so it can be used for various NLP tasks such as text classification. In the code above, the IndoBERT model is used to perform sentiment classification on the beauty product review dataset.

First, the text data consisting of reviews along with corresponding sentiment labels is split into training and validation parts using the `train_test_split` function from the scikit-learn library. This function randomly partitions the dataset into training and validation sets based on a specified ratio, ensuring that both sets represent the original data distribution. In this process, the data is typically split with an 80-20 ratio, where 80% of the data is used for training and 20% is used for validation. This stratified split helps in maintaining the balance of sentiment labels across both subsets, which is crucial for training a model that generalizes well.

Sentiment labels that are initially in text form ("Negative", "Neutral", "Positive") are converted into a numeric format for model training purposes. This conversion is necessary because machine learning models typically require numerical input. For instance, labels might be encoded as 0 for "Negative", 1 for "Neutral", and 2 for "Positive".

Next, the dataset is prepared by implementing the `ReviewDataset` class, which is a derivative of the `Dataset` class from PyTorch. This class is designed to load text and labels from the dataset and prepare them in a format that can be processed by the model. It includes tokenization using the BERT tokenizer, which converts text into a format that BERT can understand.

For training techniques, this model is trained using a fine-tuning approach, where a model that has been pretrained on a general task is re-tuned on a specific task, namely sentiment classification. This training process was carried out using a GPU, with a batch size of 16 and carried out for 3 epochs. The batch size refers to the number of samples processed by the model in each training iteration, while an epoch represents one complete pass through the entire training dataset. A smaller batch size of 16 allows for more frequent updates to the model parameters. The number of epochs determines how many times the model will see the entire dataset during training, with 3 epochs being a common choice for fine-tuning.

In addition, this training technique uses the AdamW optimization algorithm to optimize model parameters. AdamW is a variant of the Adam optimizer that includes weight decay, which helps in preventing overfitting. After the training process is complete, the model is evaluated using a validation dataset to measure the model's performance in performing sentiment classification. This evaluation provides insights into how well the model generalizes to unseen data and helps in tuning hyperparameters if necessary.The training process starts by calling the `trainer.train()` function, which will train the model based on the provided data. During training, the model parameters are adjusted iteratively using the AdamW optimization method, which has been implemented in the Transformers library. GPUs are used to speed up the training process, enabling parallel computing to increase efficiency.

The model training process in this context involves several key concepts: batch size, epoch, and iteration

- Batch size: In deep learning, batch size refers to the number of data samples processed simultaneously in one iteration during model training. Determining the batch size influences how much data is fed into the model at each step of the iteration. Different batch sizes can affect the model's convergence speed, memory usage, and the stability of the training process. Selecting the appropriate batch size can significantly impact the performance and final outcomes of the developed deep learning model [16]. In this context, a batch size of 16 means that 16 beauty product reviews are fed into the model simultaneously at each training step.. In this context, a batch size of 16 means that 16 beauty product reviews are fed into the model simultaneously at each training step

- Epoch: An epoch is a cycle of a machine learning algorithm during which the model learns from the entire training dataset, processed in mini-batches. In the context of using Convolutional Neural Networks (CNN) for classifying fashion and furniture, an epoch represents one iteration of learning for the machine. The more epochs used, the better the accuracy that can be achieved. However, there is a maximum accuracy limit that can be reached within the dataset. Using the appropriate number of epochs is crucial for achieving maximum accuracy in machine

learning [17]. In this case, the model views and learns all the beauty product reviews in the dataset 3 times (a predetermined number of epochs).

- Iteration: Iteration is a process or method used repeatedly to solve a mathematical problem. In the context of numerical methods, iteration is employed to find the numerical solution of an equation or function by repeatedly approaching from initial points until the desired solution is approximated. In programming, iteration also refers to a specific characteristic of an algorithm or computer program where a series of algorithmic steps are performed repeatedly within a program loop to achieve a particular goal[18]

Thus, the training process is carried out iteratively, where the model updates its knowledge about the data at each epoch. By the end of training, the model has better knowledge of how to classify beauty product review sentiment based on what it has learned from the training dataset.

C. Model Traning

TABLE I. TRAINING PROCESS

Running Time	
Num examples	2000
Num Epochs	3
Instantaneous batch size per device	16
Total train batch size (w. Parallel, distributed & accumulation)	16
Gradient Accumulation Steps	1
Total Optimization Steps	375

Table 1 showed the training process of this model. In this context, there are 2000 review examples used to train the model, with a total of 3 epochs. Each epoch includes iterations through the entire training dataset. The instant batch size setting per device is 16, which indicates the number of data instances processed simultaneously at each iteration by the hardware used, such as the GPU. The total training batch size, including parallel processing, distribution, and accumulation, is also 16. The training process includes 375 optimization steps, where the model is gradually fine-tuned to the training data to improve its ability to understand and predict sentiment from reviews. The results of this training are recorded in the training metrics, where the total training time is approximately 531 seconds. The training sample rate is about 11,298 examples per second, while the training steps per second is about 0.706

IV. RESULTS AND ANALYSIS

TABLE II. TRAINING PROCESS

METRIC	VALUE
Num examples	500
Batch size	16
Eval loss	0.8005783557891846
Eval accuracy	0.662
Eval runtime	15.9372 seconds
Samples/s	31.373
Steps/s	2.008
Epoch	3.0

Table 2 shows the evaluation, model evaluation was carried out using 500 test data samples with a batch size of 16. The evaluation results showed that the loss evaluation value (eval_loss) was 0.8005783557891846, while the evaluation accuracy (eval_accuracy) was 0.662. This means the model has an accuracy rate of 66.2% in predicting sentiment from test data. However, a high loss evaluation value indicates that the model still has difficulty in fitting the test data well, which leads to performance degradation.

There are several factors that can cause high evaluation loss and low accuracy. One is the lack of data representation that covers sufficient variation of the target classes in the dataset, especially if there is an imbalance between the number of samples for each class. Additionally, a complex model may suffer from overfitting to the training data, meaning it may learn patterns that are too specific to the training data to generalize well to the test data.

To improve model performance, several steps can be taken. First, expand the dataset by increasing less representative examples or by using data augmentation techniques to create additional variation in the dataset. Second, re-tuning the model architecture or re-adjusting model parameters to reduce overfitting. Additionally, techniques such as regularization or dropout can also be used to prevent overfitting. Furthermore, the evaluation of model performance can also be improved by conducting further research related to the selection of more informative features or the use of more sophisticated natural language processing techniques. By combining these various strategies, it is hoped that the model's performance can be improved in predicting sentiment more accurately on a wider test dataset.

Next, the model will be tested using a testing dataset. The model's performance on the test data will be evaluated through a confusion matrix to assess the

extent of the model's ability to accurately predict all sentiments on the test data. After obtaining the values from the confusion matrix, we can calculate accuracy, precision, recall, and F1-score. The calculation formula can be found in the following equation.

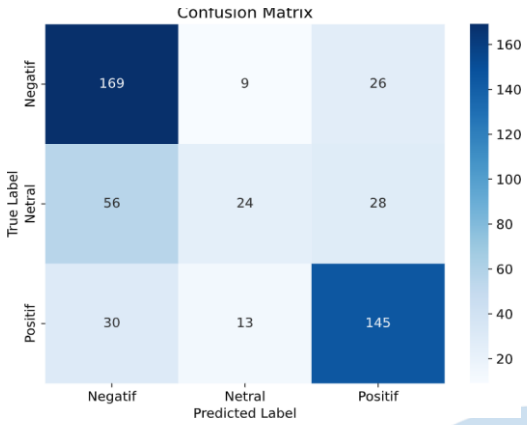

Fig. 5 Confusion Matrix

Figure 6 shows the confusion matrix, it can be seen that the model faces challenges in predicting data with neutral sentiment, but its performance is quite good for positive and negative sentiment.

Based on the results of the confusion matrix presented, it can be seen that the model has shown a relatively good level of effectiveness in conducting general sentiment analysis. From a total of 200 data samples evaluated, the model was able to accurately classify 169 cases of negative sentiment, 26 cases of neutral sentiment, and 145 cases of positive sentiment. Although there are several cases where the model misclassifies sentiment, especially in the classification between neutral and positive sentiment, the proportion of these errors is relatively small compared to the total sample evaluated.

	precision	recall	f1-score	support
Negatif	0.73	0.76	0.75	204
Netral	0.41	0.38	0.39	108
Positif	0.72	0.71	0.72	188
accuracy			0.66	500
macro avg	0.62	0.62	0.62	500
weighted avg	0.66	0.66	0.66	500

Fig. 6 Test data model evaluation

The results of model evaluation on test data as you can see on figure 7 are shown in the table above. Precision measures how accurately the model is in identifying instances that actually fall within a particular class. In this case, for negative sentiment, a precision of 0.73 indicates that of all the instances predicted as negative by the model, 73% of them are actually negative sentiment. Recall measures how well the model captures all instances that actually belong to a particular class. With a recall of 0.76 for negative sentiment, this means the model managed to find 76% of all instances that were actually negative. F1-score is the harmonic average of precision and recall, providing an overall picture of the balance between precision and recall. For negative sentiment, an F1 - score of 0.75 indicates a good balance between precision and recall. Next, support is the number of instances included in each class. Overall, the model achieved an accuracy of 0.66, which is the percentage of instances correctly predicted from the entire test data. In terms of overall evaluation, the macro avg and weighted avg values are the average of the evaluation metrics for each class, by giving the same weight or based on the amount of support for each class.

This evaluation provides an overview of the model's performance in predicting sentiment from test data, highlighting the model's strengths and weaknesses in classifying these sentiments. However, there are several notes that need to be noted. The significant difference in the number of cases misclassified between negative and positive sentiment suggests potential improvements in understanding certain linguistic nuances, such as slang or dialect, that may influence sentiment interpretation. The model may not be fully able to capture certain language variations used in the text, which can lead to errors in classification.

When evaluating a model's accuracy in classifying positive, negative, and neutral sentiment, it is necessary to consider how the model responds to text variations and complex linguistic contexts. The trained model may be able to identify common patterns in sentences that indicate certain sentiments, such as emotional words or signs of satisfaction or dissatisfaction. For example, in a restaurant review, the model may be able to recognize words such as “good” or “bad” that strongly express positive or negative sentiment. However, the model may have difficulty understanding sentences that use irony or innuendo to convey a sentiment that actually contradicts the words used.

For example, if a user writes, "Wow, the service at this restaurant was truly amazing," the model might directly interpret the word "amazing" as an indication of positive sentiment. However, in this context, the author actually uses the word ironically, because the service provided is actually bad. The model is unable to capture nuances of irony or sarcasm in text, and therefore, may produce inaccurate sentiment classifications.

It is important to remember that a model's accuracy in classifying sentiment can be affected by various factors, including language and cultural variations in the text. Languages have many nuances, idioms, and

unique ways of expressing sentiment, which may not always be easy for the model to interpret. Additionally, a model trained on a particular dataset may not be able to handle certain language or context variations well, thereby affecting its accuracy in classifying sentiment. Although the model can provide satisfactory results in sentiment classification tasks, there are limitations in its ability to understand and interpret certain linguistic nuances, such as irony or sarcasm. As a result, careful evaluation of context and language variations is important to understand the performance and limitations of models in identifying sentiment from text.

Thus, although the model has demonstrated a satisfactory level of effectiveness in conducting sentiment analysis in general, it is recommended to make adjustments or improvements to the understanding of certain slang or dialect texts to increase its accuracy and reliability in analyzing sentiment in broader and diverse contexts. Careful evaluation of a confusion matrix like this provides valuable insight into model performance and helps in identifying areas that need improvement to improve the overall quality of sentiment analysis

V. CONCLUSION

The model evaluation results show that even though the model has undergone several epochs (3.0), there are still several problems that need attention. Evaluation shows that the evaluation loss value (eval_loss) is quite high, reaching 0.8006. This shows that the model still has a significant error rate in predicting sentiment on test data. However, the evaluation accuracy (eval_accuracy) of 0.662 shows that the model succeeded in correctly predicting around 66.2% of the total examples in the test data.

The limitations seen in this evaluation may be due to several factors. One of them is the complexity of the data used. Sentiment in texts is often difficult to predict due to variations in language, context, and writing style. In addition, the presence of certain slang or dialect words in the data can also complicate the sentiment classification process. To improve model performance, future research can consider expanding the training dataset by adding more diverse and representative data. More sophisticated text processing can also be applied to address variations in language and writing style.

While there is still room for improvement, the model is still quite effective for general sentiment analysis. However, recommendations for adjustments to slang or dialect texts need to be considered so that the model can be more sensitive to variations in language. Thus, future research can lead to the development of more sophisticated models and be able to better overcome challenges in text sentiment analysis

REFERENCES

[1] Anggraini Nita and Ubaidillah Hasan, "Impact of Online Reviews, Pricing, and Viral Marketing on Shopee Purchases.," *Indonesian Journal of Law and Economics Review*, vol. 19, no. 2598–9928, pp. 10–21070, May 2024, Accessed: Sep. 04, 2024. [Online]. Available: https://ijler.umsida.ac.id/index.php/ijler/article/view/1088

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1–135, Apr. 2008, doi: 10.1561/1500000011.

[3] Md. J. Hossain, D. Das Joy, S. Das, and R. Mustafa, "Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2022, pp. 522–527. doi: 10.1109/ICISET54810.2022.9775846.

[4] B. M. D. Abighail, Fachrifansyah, M. R. Firmanda, M. S. Anggreainy, Harvianto, and Gintoro, "Sentiment Analysis E-commerce Review," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 1039–1045. doi: 10.1016/j.procs.2023.10.613.

[5] S. H. Muhammad *et al.*, "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.08956

[6] B. Liu, "Sentiment Analysis and Opinion Mining," in *Synthesis Lectures on Human Language Technologies*, Apr. 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

[7] R. Awad Mariette and Khanna, "Machine Learning," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Berkeley, CA: Apress, 2015, pp. 1–18. doi: 10.1007/978-1-4302-5990-9_1.

[8] D. Patil and N. Rane, "Customer experience and satisfaction: Importance of customer reviews and customer value on buying preference," vol. 5, pp. 3437–3447, Apr. 2023, doi: 10.56726/IRJMETS36460.

[9] V. Chandradev, I. Made, A. Dwi Suarjaya, I. Putu, and A. Bayupati, "Chandradev, Analisis Sentimen Review Hotel menggunakan Metode Deep Learning BERT 107 Analisis Sentimen Review Hotel Menggunakan Metode Deep Learning BERT," *Jurnal Buana Informatika*, vol. 14, no. 2, pp. 107–116, 2023.

[10] A. S. P. Braja and A. Kodar, "Implementasi Fine-Tuning BERT untuk Analisis Sentimen terhadap Review Aplikasi PUBG Mobile di Google Play Store," *J I M P - Jurnal Informatika Merdeka*

Pasuruan, vol. 7, no. 3, p. 120, Sep. 2023, doi: 10.51213/jimp.v7i3.779.

[11] Eza Ananda Putra, "Sentiment Analysis Using Bidirectional Encoder Representations from Transformers (BERT)," Medium. Accessed: May 03, 2024. [Online]. Available: https://medium.com/@eza.a.putra/implementasi-bert-untuk-analisis-sentimen-terhadap-ulasan-aplikasi-flip-berbahasa-indonesia-557d691e0440

[12] N. Aurelia Salsabila, U. Sa, and F. Fauzi, "Analisis Sentimen Pada Ulasan Aplikasi Tokopedia Menggunakan Klasifikasi Naïve Bayes," *PRISMA 2024*, vol. 7, pp. 44–51, 2024, [Online]. Available: https://play.google.com/store/apps/details?id=com.tokopedia.tkpd&hl=en

[13] S. Rahayu, J. Jaya Purnama, A. Hamid, and N. K. Hikmawati, "Analisis Sentimen AicoGPT (Generative Pre-trained Transformer) Menggunakan TF-IDF," *Jurnal Buana Informatika*, vol. 14, pp. 97–106, Nov. 2023.

[14] Abdul Halim Hasugian, M. Fakhriza, and Dinda Zukhoiriyah, "Analisis Sentimen Pada Review Pengguna E-Commerce Menggunakan Algoritma Naïve Bayes," *Teknologi Sistem Informasi dan Sistem Komputer TGD*, pp. 98–107, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsk/index

[15] Sekar Mulyani, "Indobertweet Ulasan Beauty Products," Hugging Face.