Khazanah Journal of Religion and Technology
Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia
https://journal.uinsgd.ac.id/kjrt
Online ISSN: 2987-6060
Volume 2 Issue 2 (2024)
DOI: https://doi.org/10.15575/kjrt.v2i2.1118

Sentiment Analysis of Marketplace Review with Islamic Perspective using Fine-Tuning DistilBERT

Reza Fahlevi Herdiyanto
Dimensi Web Community
Bandung, Indonesia
rfahlevih12@gmail.com

Rd. Imam Saepul Millah
Pergerakan Mahasiswa Islam Indonesia
Bandung, Indonesia
rdsaepulmillah@gmail.com

Muhammad Thoriq
Dimensi Web Community
Bandung, Indonesia
muhammadthoriq11@gmail.com

Received: November 20, 2024; Revised: December 20, 2024; Published: January 6, 2025

Abstract— E-commerce apps have become an important part of modern life, with user reviews playing a crucial role in assessing platform performance and identifying areas for improvement. This research aims to apply the Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) model to analyze the sentiment contained in user reviews of Tokopedia and Shopee apps on the Google Play Store. Along with the rapid growth of e-commerce in Indonesia, platforms such as Tokopedia and Shopee have become an important component in facilitating fast and easy online transactions. User reviews on these platforms are a valuable source of information for evaluating customer satisfaction, but the increasing volume of reviews makes manual sentiment analysis inefficient. This research uses a dataset of Indonesian-language reviews on Tokopedia and Shopee Apps on Google Play Store to classify sentiments into positive and negative categories by utilizing DistilBERT, which is a lightweight variant of BERT with the ability to efficiently process large data without sacrificing accuracy. The results of this analysis provide insights that can assist e-commerce platforms in improving user experience as well as support data-driven decision-making for application development. This research contributes to the application of natural language processing (NLP) technology for sentiment analysis in the context of e-commerce in Indonesia.

Keywords- DistilBERT, e-Commerce, Islamic Perspective, Sentiment Analysis, Shopee, Tokopedia

I. INTRODUCTION

E-commerce has become an essential part of everyday life in today's digital age, allowing people to shop whenever and wherever they want without any time or place restrictions. E-commerce not only helps consumers make transactions but also provides a great opportunity for businesses to expand their market reach. Based on data from Databoks, Shopee recorded the highest number of visitors, at 2.3 billion, followed by Tokopedia with 1.25 billion visitors [1].

Both platforms allow users to purchase the products and services they need easily. With the rapid growth in the number of users and the increasing number of reviews, the process of sentiment analysis of these reviews has become increasingly difficult to do manually [2]. Therefore, the use of technologies such as sentiment analysis based on natural language processing (NLP) is needed. Sentiment analysis can automatically identify whether a review is positive or negative. It provides invaluable insights to understand customer satisfaction, identify areas for improvement, and make data-driven decisions [3].

One of the deep learning models that is often used in natural language processing is Bidirectional Encoder Representations from Transformer (BERT). BERT itself has outstanding performance in various NLP tasks, including

sentiment analysis [4]. In this research, we chose to use a lighter and more efficient variant of BERT called DistilBERT. DistilBERT offers advantages in terms of speed and efficiency, enabling faster processing of large datasets without sacrificing accuracy in sentiment analysis. Despite being lighter in weight, Distil- BERT is still capable of delivering results equivalent to BERT [5]. Therefore, DistilBERT is an ideal choice for sentiment analysis in Indonesian reviews.

This research aims to utilize DistilBERT in analyzing the sentiment of user reviews of the Tokopedia and Shopee applications found on the Google Play Store. Review data from the Google Play Store is a particularly relevant source for this research, given that the platform is the primary place for users to provide feedback on their experiences with mobile apps. The large and diverse volume of reviews on Google Play Store provides both a challenge and an opportunity to apply advanced NLP techniques such as DistilBERT.

In addition to its technical significance, this research also considers the Islamic perspective on commerce and business practices. E-commerce platforms like Tokopedia and Shopee play a significant role in facilitating trade, and analyzing user reviews aligns with the Islamic principles of fairness (*adalah*), transparency (*musharahah*), and mutual satisfaction in transactions (*ridha*). By identifying areas for improvement and ensuring customer satisfaction, this research indirectly contributes to ethical commerce practices, which are foundational to Islamic teachings. Sentiment analysis in this context helps to ensure that business operations align with these values, promoting trust (*amanah*) and avoiding exploitation (*gharar*) in digital transactions.

By analyzing the sentiment in these reviews, this research hopes to provide useful insights to improve user experience on both e-commerce platforms. The results of this sentiment analysis can be used to help Tokopedia and Shopee improve customer satisfaction, identify areas for improvement, and make more informed data-driven decisions for future app development, while fostering practices that align with Islamic ethical values.

II. RELATED WORKS

Samudera, et al. addressed the need to analyze user reviews of the BSI Mobile and Action Mobile applications to understand the sentiments expressed and improve the quality of the applications. This research uses the Multinomial Naive Bayes Algorithm to classify reviews as positive or negative sentiment. The process involves several stages, including text preprocessing (case folding, tokenization, stopword removal, and stemming), division of the dataset into training and testing sets, and sentiment classification. The research dataset consists of 55,059 reviews for the BSI Mobile app and 615 reviews for the Action Mobile app, which were collected through web scraping techniques using Google's Web Scraper extension. The results show that the algorithm performs well, with 78.7% accuracy, 76.5% precision, 86.2% recall, and 80.6% f1-score for the BSI Mobile app, and 85.8% accuracy, 75% precision, 75% recall, and 75% f1-score for the Action Mobile app [6]. Wijaya, et al. examined Indonesian sentiment towards elec- tric vehicles using data from social media X (Twitter), utilizing FastText and IndoBERT algorithms for sentiment classifica- tion. The research dataset included 119,130 tweets collected between January 2020 and July 2023, with 3,000 tweets manually labeled for model training and testing. After going through six pre-processing stages, the tweets were used to train the model with various data sharing scenarios. The results showed that the IndoBERT model with a 70:30 training and testing data split scenario achieved the highest accuracy of 82.5%, outperforming the FastText model. Sentiment analysis of these tweets revealed that 58% were positive, 21.2% were negative and 20.7% were neutral. This study emphasizes the importance of public awareness of the environmental benefits of electric vehicles and the need for government efforts to encourage their adoption as part of climate change mitigation and sustainable development in Indonesia [7].

Alhadlaq, et al. discuss the challenges in aspect-based sentiment analysis (ABSA), especially in accurately classifying sentiments related to specific aspects, where traditional models often fail to understand contextual relationships and structural information in textual data. The authors propose a novel hybrid deep learning framework named DistilRoBERTa2GNN, which integrates DistilRoBERTa's ability to capture contextual relationships in text with Graph Neural Networks (GNN) to utilize structural information. The model is trained on several benchmark datasets, including Rest14, Rest15, Rest16- EN, and Rest16-ESP, which include restaurant reviews with detailed sentiment annotations. The results show strong formance, such as on the Rest14 dataset, where the model achieves an F1 score of 77.98%, with a precision of 78.12% and recall of 79.41%, demonstrating a balanced ability in capturing sentiment patterns that are nuanced and relevant to the analyzed aspects [8].

Yolanda, et al. discussed the sentiment analysis of DANA app reviews on the Google Play Store with the aim of classifying user sentiment using the Na¨ıve Bayes Classifier (NBC) algorithm. This approach includes several stages, starting from collecting review data through scraping techniques, text processing which includes cleaning, case folding, tokenization, normalization, stopword removal, and stemming, to feature selection using Information Gain (IG) to improve model efficiency and accuracy. From a dataset of 1500 Indonesian reviews, feature selection successfully reduced the number of features from 1,106 to 536, which improved model accuracy from 82.91% to 85.09%, precision from 83.96% to 85.79%, and recall from 90.23% to 92.09%. This study shows that the application of Information Gain significantly improves the performance of NBC in classifying sentiment, with the dataset divided into 80% training data and 20% testing data [9].

Agustina, et al. addressed the challenges in sentiment analysis of Shopee app reviews on Google Play Store by using Na¨ıve Bayes algorithm for classification. This research utilizes two data distribution techniques, namely Hold-Out

and K- Fold Cross Validation, to train and evaluate the model. The preprocessed review dataset uses TF-IDF method as feature weighting. With the Hold-Out method (80:20 ratio), the model achieved an accuracy of 83%, which is 1% higher than the av- erage accuracy obtained through the 10-Fold Cross Validation technique of 82%. This research demonstrates the effectiveness of the Na¨ıve Bayes algorithm with different data distributions in classifying the sentiment of Shopee app reviews [10].

Chintalapudi, et al. examined the sentiment analysis of Indian netizens during the COVID-19 pandemic using tweets to understand the emotions expressed and identify fake news related to COVID-19. This research utilizes deep learning models, specifically BERT, to classify tweets into four sentiment categories namely fear, sadness, anger, and joy. The dataset consists of 3090 tweets manually collected from the Indian Twitter platform during the COVID-19 lockdown period, i.e. from March 23, 2020 to July 15, 2020. The BERT model was trained using this dataset and showed the best performance compared to other models, with an accuracy rate of 89% in classifying the sentiment of COVID-19-related tweets. This research demonstrates the effectiveness of the BERT model in social media-based sentiment analysis tasks during the pandemic crisis [11].

Wei, et al. addressed the challenge of transferring knowledge from a full-size BERT model to a smaller model for sentiment classification tasks while maintaining high performance. The authors propose a novel learning objective that combines hidden state learning and soft label learning, denoted as Ltotal, to distill knowledge from a 12-layer BERT teacher model to a 6-layer BERT student model. The method includes refining the teacher model on the Stanford Sentiment Treebank (SST- 2) dataset, distilling knowledge using Ltotal, and refining the student model. Results show that the student model achieves 91.9% accuracy on the SST-2 development set, outperforming DistilBERT (91.3%) and maintaining about 98.2% of the performance of the full-size BERT model with a 40% reduction in parameters (from 110 million to 66 million). This research demonstrates the effectiveness of the knowledge distillation approach in generating efficient yet high-performance models [12].

Dogra et al. (2024) investigate sentiment classification of banking news events, focusing on categorizing them into three classes: positive, negative, and neutral. The authors propose using DistilBERT, a state-of-the-art deep contextual language representation model, and compare its performance with a traditional, context-independent method, TF-IDF. They fine-tune DistilBERT and test it using four supervised machine learning classifiers: Random Forest, Decision Tree, Logistic Regression, and Linear SVC. Their work highlights the benefits of deep contextual models for sentiment analysis tasks and demonstrates that DistilBERT outperforms TF-IDF across all classifiers, with Random Forest achieving 78% accuracy, a 7% improvement over TF-IDF. Furthermore, the precision and recall for all sentiment classes were higher with DistilBERT, suggesting its superiority for this task. However, the authors do not specify the exact dataset used in their experiments [13].

Joshy, et al. (2024) focus on sentiment analysis of tweets, specifically utilizing datasets related to the Coronavirus and general sentiment, such as the Sentiment 140 dataset. The authors propose using advanced transformer models—BERT, DistilBERT, and RoBERTa—to perform sentiment analysis on these datasets. Their methodology includes essential pre-processing steps, such as converting text to lowercase, removing stop words, punctuation, and special characters, as well as applying stemming and lemmatization. Tokenization is done using WordPiece tokenizers. The models are trained and tested using Keras with TensorFlow as the backend, and performance is evaluated using accuracy as the metric. For the Coronavirus tweets NLP dataset, which contains 44,955 tweets categorized into five classes (consolidated into positive and negative for analysis), and the Sentiment 140 dataset, with 1.6 million tweets divided equally into positive and negative sentiments, the BERT model outperforms the others, achieving 90.43% accuracy on the Coronavirus dataset and 93.13% on the Sentiment 140 dataset [14].

Kokab et al. (2024) address the limitations of traditional CNN-based models for sentiment analysis, particularly the issue of capturing only short-term dependency patterns due to the size of convolutional kernels, which leads to an increase in the number of parameters. To overcome these limitations, the authors propose an enhanced BERT-based CBRNN model for sentence-level sentiment analysis. This model integrates a zero-shot algorithm for data annotation, BERT for generating semantic and contextual embeddings, a dilated CNN for extracting both local and global sentimental features, and a Bi-LSTM for learning long-term dependencies. The proposed model significantly outperforms traditional CNN-based approaches, demonstrating improvements in various performance metrics, including a 0.2% increase in f1-score, 0.3% in accuracy, and 0.4% in AUC. Specifically, the model achieved high scores in recall, f-score, accuracy, and AUC, with values of 0.91%, 0.94%, 0.90%, and 0.958%, respectively, on the self-driving car dataset. It also achieved the highest precision rate of 0.98% compared to other models on the US-presidential election reviews dataset. The BERT-based CBRNN model was evaluated on multiple datasets, including the US-airline dataset (11,517 tweets), the self-driving car dataset (7,156 tweets), the US presidential election dataset (10,729 reviews), and the IMDB movie review dataset (50,000 reviews) [15].

III. RESEARCH METHODS

There are two main approaches in the development of sentiment analysis models. DistilBERT is used to create a text classification model that can find positive, and negative sentiments. Meanwhile, CRISP-DM is a development methodology that includes various stages, such as problem understanding, data preparation, model building and training, and finally result interpretation and model performance evaluation.

A. *DistilBERT*

DistilBert is a pre-trained model created using the application of knowledge distillation techniques to the larger BERT model. The architecture of DistilBERT is similar to BERT, in that both use transformers, especially the encoder block, where the transformer consists of an encoder and a decoder [16]. One of the main differences between the two is that DistilBERT eliminates pools and embeds different types of tokens. Also, the distillation model uses large pools with the help of gradient accumulation. Dynamic masking is used instead of the masking used in the original BERT and without the purpose of next sentence prediction (NSP) during training. The number of transformer (encoder) layers in the BERT base of 12 layers has been reduced to six. Therefore, the distilled version has 66 million relatively fewer parameters. The number of parameters dropped by almost half (40%), and it is faster by 60% while maintaining BERT performance of more than 95% [17]. Figure 1 provides the architecture of BERT and DistilBERT.

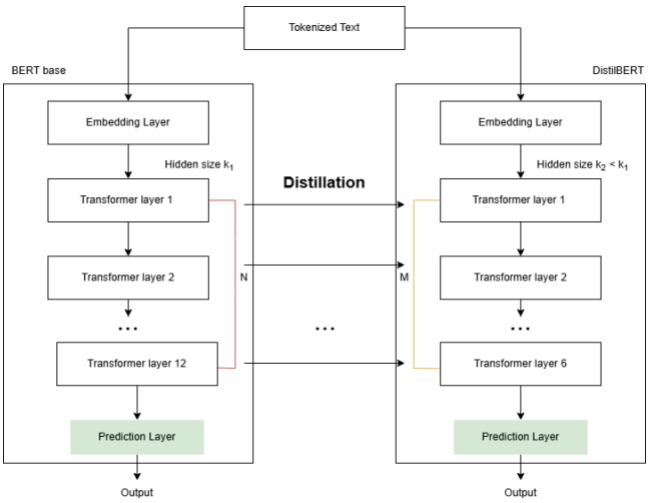

Fig. 1. DistilBERT Architecture [18]

Figure 1 illustrates how the BERT model distalization process into DistilBERT. The process aims to create a new model that is lighter and more efficient without losing too much performance. On the left, there is the BERT Base model which consists of several components, namely the Embedding Layer to translate the tokenized text into a vector representation, the Transformer Layers as many as 12 layers with a hidden size of k1, and the Prediction Layer which generates predictions based on the last representation of the transformer layer. Meanwhile, on the right is the DistilBERT model, a lighter version of BERT. DistilBERT has a similar structure but uses a hidden size of k2, which is smaller than k1, has only 6 Transformer layers, as well as a Prediction Layer to produce a prediction output comparable to the BERT Base model.

The distillation process is at the core of this diagram, where DistilBERT is trained using knowledge from BERT Base through a teacher-student technique. The smaller model (DistilBERT) is guided to mimic the results of the large model (BERT Base) by transferring important information from the transformer layer of BERT Base to the DistilBERT layer. This process often uses additional loss functions, such as soft loss, to improve the learning quality of the DistilBERT model. Distillation aims to improve efficiency by making the model faster in inference as it only uses six layers versus 12 layers in BERT Base, as well as reducing the model size and computational requirements while retaining most of BERT's language understanding capabilities. DistilBERT models are often used for applications that require high speed and low memory, such as real-time inference in resource-constrained devices.

DistilBERT is faster and smaller than BERT. This shows that the BERT base model automatically acquires labels and inputs from text, even though it was previously trained only with raw text without human labels. Therefore, this model can use a lot of publicly available data [18].

B. *CRISP-DM*

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a methodology used to manage data mining projects systematically and efficiently. CRISP-DM has been the most common data mining standard since it was first introduced in 2000 and has become the most popular standard in the industry. The main advantage of CRISP-DM is that it is industry independent, and can be applied across industries, allowing changes during the project [19]. CRISP-DM is used as a framework for systematically managing projects. This methodology allows us to focus on selecting the right model architecture. With CRISP-DM's iterative approach, we can make adjustments to the deep learning model, such as optimizing parameters, improving data quality, or modifying the neural network architecture to achieve the best performance according to the research objectives. In Figure 2 below, are the stages contained in the CRISP-DM methodology, as follows:

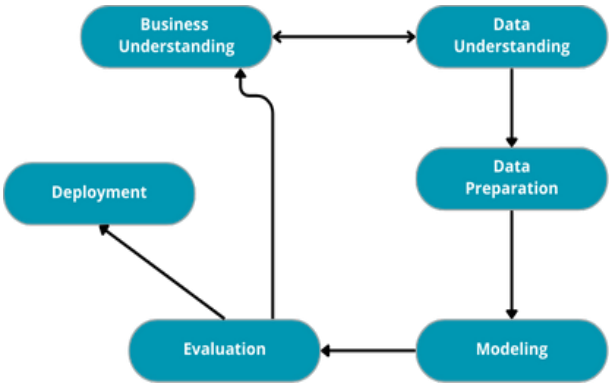

Fig. 2. CRISP-DM Methodology [20]

The CRISP-DM methodology starts with the Business Understanding stage and ends at the Deployment stage. Further explanation of these stages is as follows:

1) Business Understanding

The first step is to understand the purpose of this project. The goal may be prediction, grouping, or other specific tasks. For the next steps to be clear, the problem must be clearly described.

2) Data Understanding
At this stage, data is collected and examined to understand its meaning. We look at patterns, distributions, and possible issues such as missing or inconsistent data. From this we can tell if the data is sufficient and suitable for training the model.
3) Data Preparation
The data is divided into data for training, validation, and testing, and processed to make it ready for use, such as cleaning the data from errors, changing its format, or making it more uniform. This is done so that the model can learn well and show fair results when tested.
4) Modeling
At this stage, the algorithm or model type that best suits the problem at hand is selected, and the model is trained using the training data. In addition, its parameter settings are optimized. Experiments are conducted to find the most effective combination.
5) Evaluation
Once the model is trained, the results are tested to ensure that its performance meets expectations. We know how well the model works by using metrics such as accuracy or F1 score. If the results are unsatisfactory, we can go back to the previous stage to improve it.
6) Deployment
Tested models can be applied to usable systems, such as applications, APIs, or presented in the form of reports. This stage ensures that the results of the model can produce actual benefits.

IV. RESULT AND DISCUSSION

A. *Analysis and Result*

In this section, the results of the research are presented, including data, findings, and information relevant to the search objectives. This research uses the CRISP-DM method, which starts with business understanding and data understanding, followed by data preparation, modeling, evaluation, and model application. The CRISP-DM methodology was used systematically in this project, progressing through the various stages outlined in the framework.

By ensuring a structured approach to data mining and deep learning tasks, each stage contributed to the overall success of the project. The comprehensive and structured approach of the CRISP-DM methodology enabled the project to address a wide range of issues, such as data cleaning, model selection, and hyperparameter optimization. Iterations performed at each step resulted in an excellent DistilBERT-based final model for Indonesian text, with high accuracy and F1 scores. This makes it a reliable tool for user sentiment analysis. The following is an explanation of the stages carried out in the application of the CRISP-DM method in this research:

1) Business Understanding

The main objective of this research is to analyze the sentiment of user reviews of the Tokopedia and Shopee applications on Google Playstore. This analysis aims to understand the level of user satisfaction and identify services that are considered poor. These two apps are the largest and most widely used e-commerce platforms in Indonesia. Some of the main questions that are the focus of this analysis include understanding overall user sentiment, identifying themes and issues that often appear in reviews, and comparing positive and negative sentiment on Tokopedia and Shopee reviews. With this sentiment analysis, it is expected to better understand the user experience and address issues to improve the quality of existing services.

2) Data Understanding

The data used in this research is taken from the Tokopedia and Shopee e-commerce platforms on Google Playstore. The data consists of user reviews, ratings (from 1 to 5 stars), and timestamps. The amount of data used is 10,000 reviews, consisting of 5,000 reviews from Tokopedia and 5,000 reviews from Shopee, which are then used as datasets. The data was collected in the period from November 11, 2024, to December 11, 2024. The visualization of the frequency of each rating level can be seen in Figure 3.

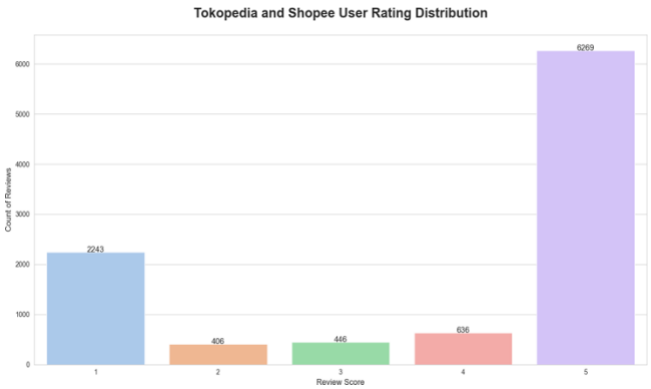

Fig. 3. Rating Distribution

We can see that the rating distribution visualization provides an overview of how user reviews and ratings are distributed on both platforms. These ratings also reflect the user's level of satisfaction with the product or service purchased. The next analysis was to calculate the number of reviews data for each week. This visualization is useful to show the trend of changes in reviews over time, as well as provide insight into the pattern of user activity in providing reviews during the observed period. The visualization can be seen in figure 4 below.

In Figure 4 of the data visualization above, the data captured covers a period of 5 weeks. Each week is accompanied by the date of the period as well as the number of reviews received in that week. In the last two weeks, we can

see an increase in the number of reviews of more than 1,000 reviews each week. The last analysis performed was the calculation of the frequency of the most frequently occurring words. In this analysis, we selected 15 words that appeared most frequently in the reviews. In Figure 5, we see a list of the most frequently occurring words.

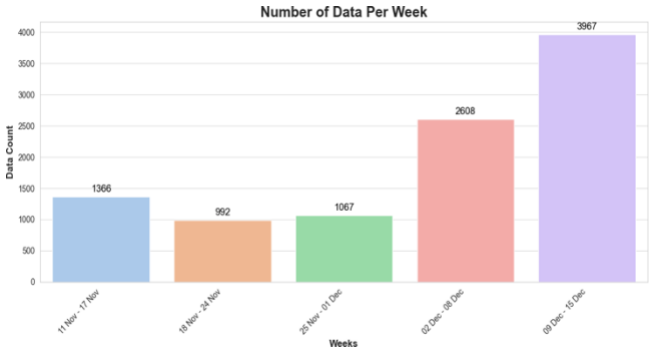

Fig. 4. Number Of Data Per Week

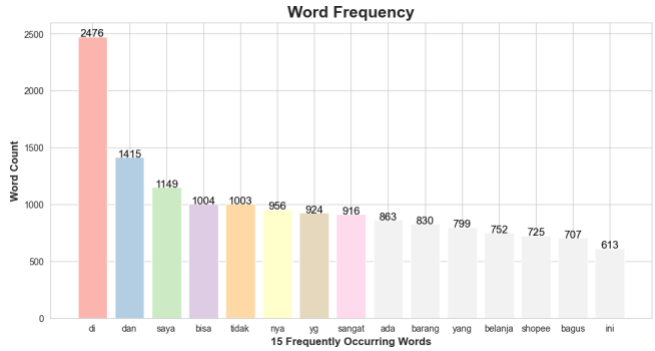

Fig. 5. Word Frequency

The analysis in Figure 5 above provides insight into the most common terms used by users, which gives an initial indication of the themes that are frequently discussed. This initial exploration is very important as it helps to understand the context and content of the reviews, which forms the basis for further analysis at a later stage. In addition, the results of this analysis can provide an overview of trends and topics that attract users' attention.

3) Data Preparation

In the data preparation stage, several steps are taken from data processing to data cleaning. First of all, we will process the data and remove irrelevant columns, such as userName, reviewId, userImage, and so on. After that, we check if there are any missing values or duplicate data. If any missing value or duplication is found, the data will be deleted immediately. Next, in the Data Preprocessing stage, we will carry out several steps that can be seen in the figure below.

The data preprocessing steps taken can be seen in the figure above, starting with changing letters to lowercase, removing emojis, removing numbers, removing punctuation, and removing whitespaces. Next, we normalize the text by replacing slang words into words that are in accordance with standard Indonesian language rules. We also removed stopwords, stemming, labeled the data, and tokenization. Overall, these steps ensure that the data is clean and ready for use.

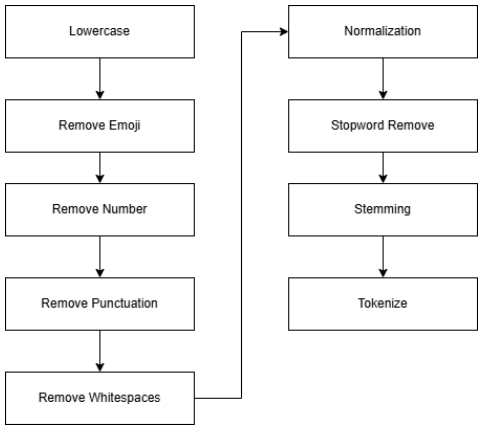

Fig. 6. Data Preprocessing Step

The most time-consuming step is stemming, which takes about 15 minutes. In the first stage, data was divided to separate the labeled and unlabeled data. The data is divided with a proportion of 80% for labeled data and 20% for unlabeled data, which are called data_labeled and data_unlabeled, respectively. The data_labeled amounted to 8,000 data, while the data_unlabeled amounted to 2,000 data. Next, for data_labeled (8,000 data), we further divided it into two parts: 80% is used for training and 20% for validation. As for the data_unlabeled (2,000 data), we did not perform labeling and left this data for model testing in the final stage. The trained model was then tested to assess its ability to guess sentiment from the unlabeled data. Tokenization is done to break the text into smaller units (usually words or sub-words), so that the model can process and analyze the text data more effectively. Overall, these preprocessing steps ensure that the data is clean and in a suitable format to be fed into the DistilBERT model.

4) Modeling

In the modeling stage, we chose the DistilBERT-based model of cahya/distilbert-base-indonesian available on Hugging Face for sentiment analysis. We chose this model after considering that DistilBERT is more lightweight compared to other DistilBERT models. The model will be trained using pre-prepared data. Hyper-parameter adjustment is highly considered during this process. this process. We performed several iterations to find the optimal configuration.

In the initial experiments, we encountered an overfitting problem which might be caused by the small amount of data and inappropriate parameter settings. In that experiment, the model could not reach more than 5 epochs because it kept overfitting. However, in the next experiment, we adjusted the learning_rate and increased the batch_size. After some

consideration, we decided to use 8 epochs. The results of the obtained model are as in Table 1.

Table 1 Training And Validation Performance

Epoch	Training Loss	Validation Loss	Accuracy	F-1
1	0.479900	0.403610	0.833755	0.885403
2	0.359600	0.325432	0.867257	0.905405
3	0.293100	0.294804	0.877370	0.912058
4	0.280100	0.280712	0.889381	0.919540
5	0.272100	0.276206	0.891277	0.921245
6	0.283000	0.274940	0.891277	0.921389
7	0.270400	0.273735	0.892541	0.922303
8	0.274700	0.273606	0.892541	0.922303

As can be seen in the table above, during training, the model shows consistent performance improvement, both in terms of training loss and validation loss, which reflects the model's ability to gradually reduce the pre- diction error. In addition, there is also a visualization of the training stage of this model as shown below:

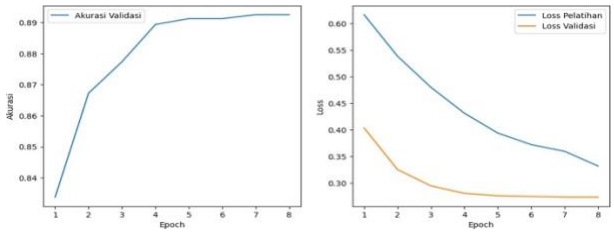

Fig. 7. Training and Validation Performance

From the graph above, we can see that train_loss and validation_loss both decrease with each epoch, while validation_accuracy continues to increase with each epoch. Therefore, it can be concluded that the model learns well, with the accuracy and F1 value indicating the model's ability to classify sentiment very well.

5) Evaluation

After the modeling stage, we proceed to the evaluation stage. At this stage, we will evaluate the previously created model using accuracy and F1 Score metrics. Based on the evaluation results, the 8th epoch shows the best performance compared to other epochs. This can be seen from the lowest validation loss of 0.273606, which indicates a good generalization ability of the model to data that has never been seen before. Although the training loss at the 8th epoch is slightly higher at 0.274700, the value is quite close to the lowest value, which indicates that the model does not experience overfitting and is able to learn the training data well. In addition, the accuracy and F1 score at the 8th epoch reached 89.25% and 92.23% respectively, which is the highest value among other epochs. These high values of accuracy and F1 score indicate that the model can classify the data well and can overcome the data imbalance problem effectively. Overall, the 8th epoch shows the best balance between validation loss, accuracy, and F1 score, so it can be considered as the most optimal epoch among all tested epochs.

In addition, we also use the confusion matrix to evaluate the performance of the classification model by comparing the model's predictions with the actual labels. The visualization of the confusion matrix can be seen in the Figure 8.

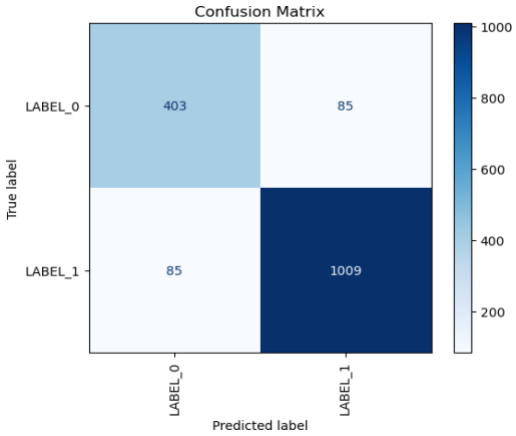

Fig. 8. Confusion Matrix

In the confusion matrix image shown, True Positives (TP) refers to the number of reviews that are actually positive and predicted as positive by the model, with a value of 1009. True Negatives (TN) is the number of reviews that are actually negative and predicted as negative by the model, with a value of 403. False Positives (FP) is the number of reviews that are actually negative but predicted as positive by the model, with a value of 85. False Negatives (FN) is the number of reviews that are actually positive but predicted as negative by the model, with a value of 85.

It can be concluded that the model managed to classify positive and negative reviews well, although there were some prediction errors, namely 85 False Positives and 85 False Negatives. In addition, we also evaluated the performance of the model on data_unlabeled. On approximately 2000 data_unlabeled, the model was reused to predict the data with the result available in Figure 9.

```
def is_mismatch(row):
    if row['score'] in [4, 5] and row['predicted_label'] != 1:
        return True
    elif row['score'] in [1, 2, 3] and row['predicted_label'] != 0:
        return True
    return False

mismatch_data = df_predict_label[df_predict_label.apply(is_mismatch, axis=1)]

print(f"Number of incorrect predictions: {len(mismatch_data)}")
```

```
Number of incorrect predictions: 198
```

Fig. 9. Wrong Predictions

It can be seen that the model performs very well in predicting. This is evident from the figure above, where the model only has errors on 198 data out of a total of about 2,000 data, which reflects the excellent accuracy and effectiveness of the model in handling review data. The previously predicted data was combined with the data used for model training. After merging, the amount of data is approximately 9,700 data. From this data, visualizations were made for the sentiment proportion and word cloud of each label. Each visualization can be seen as follows:

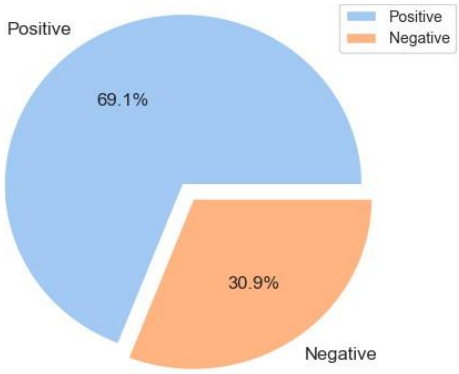

Fig. 10. Sentiment Distribution Visualization

The figure above shows that the positive sentiment is 69.1%, while the negative sentiment is only 30.1%. This shows that most of the reviews in the dataset tend to be positive, which indicates that consumers are sharing more good experiences about the product or service. Trends like this can be a strong indicator of customer satisfaction levels, which are usually related to the quality and effectiveness of the product or service. Positive reviews can be a driving factor for brand loyalty, repeat purchases, and word-of-mouth recommendations, all of which support a product or service's success in the market.

In addition, the predominance of positive sentiment can also indicate that the product or service has met or even exceeded customer expectations, which in turn strengthens the company's reputation and can improve its market position. It could also reflect a successful marketing strategy, good customer service, or a quality product. However, keep in mind that the number of positive reviews can also be influenced by biases in the request for reviews or the demographic characteristics of the customers providing the reviews.

Fig. 11. Word Cloud for Positive Sentiment

In the Figure 11, there is a word cloud for positive sentiment that illustrates the words that appear most frequently in positive reviews. The analysis shows that the dominant words in the reviews reflect consumer satisfaction with the product or service. Some of the most frequently used terms include *bagus* (good), *suka* (like), *terbaik* (best), *cepat* (fast), *memuaskan* (satisfying), *rekomendasi* (recommended) and *kualitas* (qualified).

Fig. 12. Word Cloud for Negative Sentiment

In the figure above, there is a wordcloud for negative sentiment that illustrates the words that appear most frequently in negative reviews. The analysis shows that the dominant words in the reviews reflect These words generally reflect dissatisfaction or complaints from consumers about the product or service. Some of the most frequently used terms include *buruk* (bad), *lambat* (slow), *kecewa* (disappointed), *rusak* (damaged), *gagal* (failed), and *complain* (complained).

6) Deployment

The last stage in the development process using the CRISP-DM method is deployment. In this research, deployment is done by deploying the trained model to the Hugging Face platform. This model can be accessed and used by the wider community. Figure 13 is documentation that the model is already on Hugging Face.

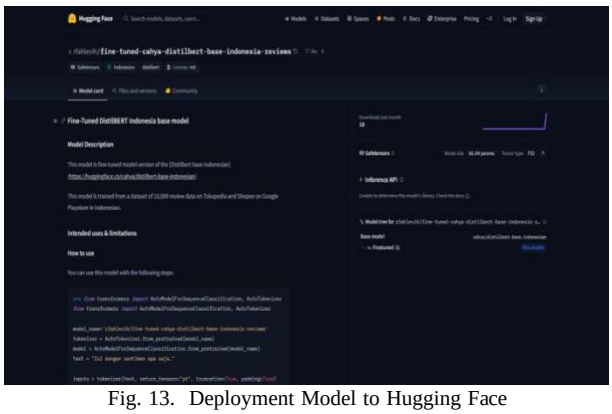

Fig. 13. Deployment Model to Hugging Face

In addition to deploying the model to Hugging Face, we also developed a simple dashboard using Streamlit. This dashboard has two main features, namely a dashboard of research results and a user sentiment-checking tool.

a) Research Result Dashboard

This dashboard displays various results of our research, including data visualization and analysis that has been done. Here is a view of the research results dashboard:

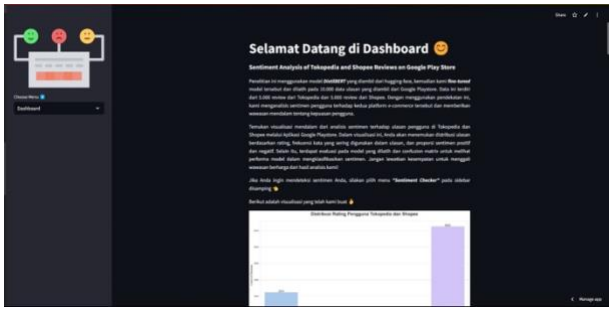

Fig. 14. Dashboard Streamlit

b) User Sentiment Checking

This feature allows users to check the sentiment of the entered text, whether it is positive or negative. Here is a view of the sentiment-checking tool:

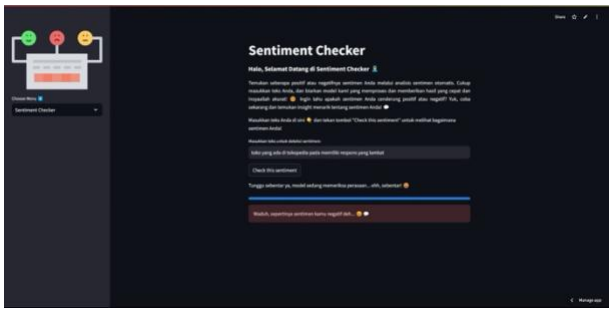

Fig. 15. User Sentiment Checking

With these two deployment steps, we hope to provide broad benefits to the community and make it easier to use the model we have developed.

B. Discussion

The results of this research demonstrate that the DistilBERT model can effectively classify user review sentiments on Tokopedia and Shopee with high accuracy (89.25%) and an impressive F1-Score (92.23%). This capability to analyze and categorize user sentiments supports the principle of *ihsan* (excellence) in Islamic commerce, which encourages continuous improvement in trade practices to ensure customer satisfaction [21]. By identifying both positive and negative aspects of user experiences, this research provides actionable insights for e-commerce platforms to enhance their services, thus fostering fairness (*adalah*) and transparency (*musharahah*) in their operations [22].

The sentiment distribution reveals that 69.1% of the reviews are positive, with keywords such as "good," "fast," and "satisfactory" reflecting customer approval of speed and product quality. This aligns with Islamic teachings that emphasize fulfilling promises and providing quality goods and services, as stated in the Quran: "*And fulfill the covenant. Indeed, the covenant is ever [that about which one will be] questioned.*" (Surah Al-Isra, 17:34). Ensuring reliable and prompt service embodies the Islamic value of trust (*amanah*) [23], which is critical in fostering strong customer relationships.

Conversely, 30.1% of negative reviews highlight complaints about slow delivery and poor product quality, with keywords like "slow," "bad," and "disappointed." From an Islamic perspective, addressing these issues is essential to avoid dissatisfaction (*gharar*) and potential harm (*dharar*) in trade [24]. E-commerce platforms can use this feedback to align their practices with Islamic principles by improving service delivery and ensuring product integrity. These efforts contribute to the concept of mutual satisfaction (*ridha*), which is central to ethical commerce in Islam [25], [26]. The limitations of this study, such as the restricted data source and class imbalance, also have implications from an Islamic perspective. Addressing these limitations in future research reflects the principle of continuous self-improvement (*islah*), which is encouraged in Islam.

By expanding data sources, balancing sentiment distribution, and employing advanced modeling techniques, future studies can contribute more comprehensively to fostering ethical practices in e-commerce. Overall, the findings of this research support the application of artificial intelligence in enhancing customer satisfaction while promoting Islamic ethical values in digital commerce. By leveraging technology responsibly and addressing user concerns, platforms like Tokopedia and Shopee can ensure that their operations align with the principles of fairness, transparency, and trust that Islam advocates.

V. CONCLUSION

This research demonstrates the effectiveness of DistilBERT in analyzing Indonesian-language user reviews, achieving an accuracy of 89.25% and an F1-Score of 92.23%. The results highlight the model's ability to generalize well and efficiently process large datasets. Sentiment distribution revealed that 69.1% of reviews were positive, indicating user satisfaction with speed and product quality, while 30.1% of reviews were negative, highlighting issues such as slow delivery and poor product quality. From an Islamic perspective, this study contributes to promoting ethical business practices, emphasizing principles like fairness (*adalah*), transparency (*musharahah*), and trust (*amanah*), while leveraging artificial intelligence to enhance customer satisfaction.

Future research should focus on expanding data sources to include reviews from multiple platforms, addressing class imbalance using advanced techniques, and incorporating multilingual datasets to provide broader insights. As a recommendation, future research can expand the data coverage by involving more diverse review sources and longer data collection periods. In addition, model development can consider other techniques to handle class imbalance and improve prediction accuracy on more complex datasets. Additionally, integrating Islamic ethical principles into AI models can offer tailored sentiment analysis for businesses in

predominantly Muslim contexts. Exploring real-time sentiment analysis and alternative transformer-based models could further enhance performance and usability. By addressing these areas, future studies can advance the application of artificial intelligence in e-commerce while promoting ethical practices that align with Islamic values.

REFERENCES

[1] A. Ahdian, "5 E-Commerce dengan Pengunjung Terbanyak Sepanjang 2023," Databoks. Accessed: Nov. 25, 2024. [Online]. Available: https://databoks.katadata.co.id/teknologitelekomunikasi/statistik/3c9132bd3836eff/5-e-commerce%20dengan-%20pengunjung-terbanyak-sepanjang-2023

[2] L. Bharadwaj, "Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification," *International Journal For Multidisciplinary Research*, vol. 5, no. 5, Sep. 2023, doi: 10.36948/ijfmr.2023.v05i05.6090.

[3] Oluwatosin Abdul-Azeez, Alexsandra Ogadimma Ihechere, and Courage Idemudia, "Enhancing business performance: The role of data-driven analytics in strategic decision-making," *International Journal of Management & Entrepreneurship Research*, vol. 6, no. 7, pp. 2066–2081, Jul. 2024, doi: 10.51594/ijmer.v6i7.1257.

[4] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based BERT model in sentiment analysis," *Applied and Computational Engineering*, vol. 67, no. 1, pp. 280–286, May 2024, doi: 10.54254/2755-2721/67/2024MA.

[5] R. Silva Barbon and A. T. Akabane, "Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study," *Sensors*, vol. 22, no. 21, p. 8184, Oct. 2022, doi: 10.3390/s22218184.

[6] B. D. Samudera, N. Nurdin, and H. A. K. Aidilof, "Sentiment Analysis of User Reviews on BSI Mobile and Action Mobile Applications on the Google Play Store Using Multinomial Naive Bayes Algorithm," *International Journal of Engineering, Science and Information Technology*, vol. 4, no. 4, pp. 101–112, Oct. 2024, doi: 10.52088/ijesty.v4i4.581.

[7] D. R. Wijaya, G. M. A. Sasmitha, and W. O. Vihikan, "Sentiment Analysis of Indonesian Citizens on Electric Vehicle Using FastText and BERT Method," *Journal of Information Systems and Informatics*, vol. 6, no. 3, pp. 1360–1372, Sep. 2024, doi: 10.51519/journalisi.v6i3.784.

[8] A. Alhadlaq and A. Altheneyan, "Distilroberta2gnn: a new hybrid deep learning approach for aspect-based sentiment analysis," *PeerJ Comput Sci*, vol. 10, p. e2267, Aug. 2024, doi: 10.7717/peerj-cs.2267.

[9] Cindy Caterine Yolanda, Syafriandi Syafriandi, Yenni Kurniawati, and Dina Fitria, "Sentiment Analysis of DANA Application Reviews on Google Play Store Using Naïve Bayes Classifier Algorithm Based on Information Gain," *UNP Journal of Statistics and Data Science*, vol. 2, no. 1, pp. 48–55, Feb. 2024, doi: 10.24036/ujsds/vol2-iss1/147.

[10] N. Agustina, D. H. Citra, W. Purnama, C. Nisa, and A. R. Kurnia, "Implementasi Algoritma Naive Bayes untuk Analisis Sentimen Ulasan Shopee pada Google Play Store," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 2, no. 1, pp. 47–54, Apr. 2022, doi: 10.57152/malcom.v2i1.195.

[11] N. Chintalapudi, G. Battineni, and F. Amenta, "Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models," *Infect Dis Rep*, vol. 13, no. 2, pp. 329–339, Apr. 2021, doi: 10.3390/idr13020032.

[12] S. Wei, D. Yu, and C. Lv, "A Distilled BERT with Hidden State and Soft Label Learning for Sentiment Classification," *J Phys Conf Ser*, vol. 1693, no. 1, p. 012076, Dec. 2020, doi: 10.1088/1742-6596/1693/1/012076.

[13] S.-L. Peng, S.-Y. Hsieh, S. Gopalakrishnan, and B. Duraisamy, *Intelligent Computing and Innovation on Data Science*, vol. 248. Singapore: Springer Nature Singapore, 2021. doi: 10.1007/978-981-16-3153-5.

[14] A. Joshy and S. Sundar, "Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa," in *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*, IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/IPRECON55716.2022.10059542.

[15] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, p. 100157, Jul. 2022, doi: 10.1016/j.array.2022.100157.

[16] A. Areshey and H. Mathkour, "Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet," *Expert Syst*, vol. 41, no. 11, Nov. 2024, doi: 10.1111/exsy.13701.

[17] S. Ikiss, N. Daoudi, M. Abourezq, and M. Bellafkih, "Exploring the potential of DistilBERT architecture for automatic essay scoring task," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 2, p. 1234, Nov. 2024, doi: 10.11591/ijeecs.v36.i2.pp1234-1241.

[18] S. V *et al.*, "The DistilBERT Model: A Promising Approach to Improve Machine Reading Comprehension Models," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 8, pp. 293–309, Sep. 2023, doi: 10.17762/ijritcc.v11i8.7957.

[19] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[20] U. Kannengiesser and J. S. Gero, "Modelling the Design of Models: An Example Using CRISP-DM," *Proceedings of the Design Society*, vol. 3, pp. 2705–2714, Jul. 2023, doi: 10.1017/pds.2023.271.

[21] M. Muhammad and M. R. Muhammad, "Building trust in e-commerce: A proposed shari'ah compliant model," *Journal of Internet Banking and Commerce*, vol. 18, Dec. 2013.

[22] M. Akram, N. Khan, and M. N. Anjum, "Perceived Financial Transparency and Loyalty in the Islamic Banking Sector of Pakistan: Exploring the Role of Trust and Age," *Contemporary Issues in Social Sciences and Management Practices*, vol. 3, no. 2, pp. 14–25, Jun. 2024, doi: 10.61503/cissmp.v3i2.160.

[23] M. A. Khan, "Justice in economics: an Islamic perspective," in *Islamic Economics and Human Well-being*, Edward Elgar Publishing, 2024, pp. 66–108. doi: 10.4337/9781035333691.00012.

[24] M. Roberts-Lombard and D. J. Petzer, "Do you want my loyalty? Then understand what drives my trust – a conventional and Islamic banking perspective," *International Journal of Islamic and Middle Eastern Finance and Management*, vol. 17, no. 3, pp. 532–551, Jul. 2024, doi: 10.1108/IMEFM-10-2023-0412.

[25] E. Elmahjub, "Artificial Intelligence (AI) in Islamic Ethics: Towards Pluralist Ethical Benchmarking for AI," *Philos Technol*, vol. 36, no. 4, p. 73, Dec. 2023, doi: 10.1007/s13347-023-00668-x.

[26] K. Albar, A. Abubakar, and A. Arsyad, "Islamic Business Ethics in Online Commerce: A Perspective from Maqashid Shariah by Imam Haramain," vol. 07, no. 2, pp. 273–289, 2023, doi: 10.33852/jurnalin.v7i2.501.