J-COSINE (Journal of Computer Science and Informatics Engineering)
Vol. 9, No. 2, December 2025 E-ISSN:2541-0806
Accredited Sinta-4 by RISTEKDIKTI Decree No. 79/E/KPT/2023 P-ISSN:2540-8895

Multitask Aspect-Based Sentiment Analysis of Indonesian Tweets on Mandalika Circuit using CNN and IndoBERTweet Embeddings

Raissa Calista Salsabila*, Ramaditia Dwiyansaputra, I Gede Pasek Suta Wijaya
Department of Informatics Engineering, University of Mataram
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: calistaahusni@gmail.com, [rama, gpsutawijaya]@unram.ac.id

Corresponding Author

Abstract. This study proposes a multitask Aspect-Based Sentiment Analysis (ABSA) model for Indonesian tweets related to the Mandalika Circuit, using IndoBERTweet embeddings and Convolutional Neural Networks (CNN). The model simultaneously predicts aspect categories and sentiment polarities. Two experimental setups were evaluated: one using raw tweets (Scenario 1) and another with preprocessed text (Scenario 2). The results show that Scenario 1 consistently outperforms Scenario 2, highlighting the ability of IndoBERTweet to handle informal tweet structures without requiring standard text cleaning. A paired t-test was conducted to evaluate statistical differences in performance between scenarios. While Scenario 1 showed higher average F1-scores, the p-value (0.7178) suggests no statistically significant improvement across all classes. Further analysis reveals that certain classes, primarily neutral and positive sentiments, tend to perform worse than negative sentiments. Data augmentation was shown to improve recall and help the model handle underrepresented classes, particularly for "Ekonomi-Negative" and "Fasilitas-Negative" labels. The study highlights the importance of preserving informal language structures and utilizing data augmentation to enhance ABSA performance on real-world tweet data.

Key words*: *Aspect-Based Sentiment Analysis, Convolutional Neural Network, IndoBERTweet, Mandalika Circuit, Multi-Task Learning

I. INTRODUCTION

The Mandalika International Circuit is situated in the Lombok Tengah district, Nusa Tenggara Barat (NTB), and is part of the Mandalika Special Economic Zone's development. The construction of the Mandalika International Circuit is designed to accelerate economic growth and development in Lombok [1]. From 2018 to 2023, tourist visits to Mandalika have increased at an average annual rate of 53.7%, including both foreign and domestic visitors [2]. Despite the growth, the construction of the circuit continued to face protests concerning land disputes [1], community exclusion from the development process, and the social-economy effect [3]. These mixed public reactions underscore the importance of understanding how people respond to various aspects of the project. They offer valuable insights into public opinion and social impact.

X (formerly known as Twitter) is a widely used social media platform with millions of users worldwide, generating real-time public discourse [4]. According to [5] Indonesia had 24.85 million active users, placing the country fifth globally in terms of active user count. Given its popularity and dynamic content, numerous studies have explored the use of X as a resource for developing predictive systems, including in public sentiment analysis across various domains [4].

Sentiment analysis is a Natural Language Processing (NLP) task that predicts sentiment towards a single topic within a text. In recent years, this task has developed to fulfill the need to recognize more fine-grained aspect-level opinions and sentiments, known as Aspect-based Sentiment Analysis (ABSA). ABSA focuses on the target of the sentiment, shifting from an entire sentence or document to an entity or a specific aspect of an entity [6]. For example, in a sentence "Makanan ini lezat" (this food is delicious). It is desired to extract the aspect terms "makanan" (food) along with the positive sentiment polarity indicated by the phrase "lezat" (delicious).

Many studies have developed ABSA, and in recent years, it has evolved into more advanced approaches such as multi-task learning. This framework allows the model to predict aspect categories and sentiment polarities simultaneously, improving their ability to understand the common features between the aspect and the sentiment [7]. Therefore, an appropriate approach is required to extract such features from the text effectively, and a convolutional neural network (CNN) is one widely used method [8].

CNN is particularly well-suited for ABSA due to its ability to capture local patterns in text by applying filters that detect keyword patterns within sentences. However, while CNN is effective at feature extraction, it requires a large amount of training data and struggles to capture long-range context in sentences [8].

To address these limitations, recent advancements have led to the development of contextual word embeddings designed for specific languages. In Indonesia, one such model is IndoBERTweet. IndoBERTweet is a domain-adapted version of IndoBERT that has been fine-tuned on Indonesian Twitter data. Its architecture makes it

well-suited for processing noisy, user-generated content typically found on platforms like X [9].

A previous study by [10] investigates multi-task ABSA for the Mandalika Circuit using FastText embeddings combined with a CNN architecture. While their approach has shown promising results, FastText generates static word representations and does not capture contextual nuances, which can be crucial in social media text where word meanings often depend on surrounding context. To build upon this prior work and ensure fair comparison, we evaluate our proposed IndoBERTweet-CNN not only on our newly constructed dataset but also on the dataset used in the previous study.

Building upon this, this study aims to develop a multi-task framework with CNN and IndoBERTweets as the embedding layers for aspect-based sentiment analysis. In this case, ABSA can be specifically beneficial in understanding how the public perceives government policy, its effects on tourism, the economy, politics, the MotoGP event, and other emerging issues related to the Mandalika International Circuit.

II. Related Works

Previous research on ABSA in the Indonesian language has covered domains such as reviews on restaurants [11], tourist attractions [12], hotels [13], and e-commerce products [14], as well as political discourse on X [15]. While these studies demonstrate the feasibility of aspect and sentiment analysis in the Indonesian language, many have treated aspect detection and sentiment classification as separate tasks, limiting their ability to model the relationship between aspects and sentiments effectively.

Multi-task learning offers a solution to this limitation by enabling models to jointly predict aspects and sentiments, capturing shared features and improving generalization. Multi-task learning has been widely applied in ABSA research, particularly in benchmark datasets such as SemEval [16], [17], [18]. These studies demonstrate that sharing parameters across tasks enables the model to better capture the relationship between opinion targets and sentiments, thereby improving accuracy and generalization. Recent efforts have also begun to apply multi-task learning in Indonesian ABSA tasks [18], [10]. Study [18] evaluated IndoBERT models with and without multi-task learning for ABSA analysis. Their multi-task learning approach achieved higher performance, with an F1-score of 96.16% compared to 94.78% for the non-MTL model. However, their work focused on structured review rather than informal social media text, and fine-tuning large transformer models like IndoBERT remains computationally intensive.

The study [10] conducted the most related approach with a similar study case to ours. The proposed work introduced an approach utilizing a CNN-based architecture with FastText embeddings to jointly predict aspect and sentiment classes, achieving accuracies of 84% for aspect classification and 72% for sentiment classification. In terms of architecture, many other ABSA studies have been conducted using different approaches, including recurrent neural networks (RNN) and fine-tuning a transformer-based model. The RNN, including LSTM and bi-LSTM variants, processes sequences one step at a time and is effective in modeling long-range dependencies in text. However, they tend to be slower and may overlook crucial local patterns in shorter texts, such as tweets [19]. In contrast, transformer-based models capture context across entire sentences simultaneously through a self-attention mechanism, achieving strong performance when fine-tuned for specific tasks. Yet, fine-tuning transformers is computationally demanding and can lead to overfitting on smaller datasets [20]. CNN, meanwhile, are well-suited for detecting local patterns by applying filters over small windows of text, capturing n-gram features that are relevant in short text. They process data in parallel, making them computationally efficient compared to sequential models like RNN. However, CNN may struggle to capture long-range dependencies across entire sentences, which can limit their ability to fully understand context when sentiments or aspects depend on broader discourse [8].

To overcome this limitation, contextual embeddings have been introduced. Unlike statistical word embeddings such as FastText, which generate the same vector for a word regardless of its surrounding context, contextual embeddings create word representations that incorporate information from surrounding words. This enables models to better capture nuanced meanings and long-range dependencies in text. In Indonesian, IndoBERTweet has been developed to handle informal and noisy data better. IndoBERTweet builds on IndoBERT by continuing its training with a large-scale Indonesian Twitter dataset of 26 million tweets from various domains. This approach allows the model to adapt to the informal language commonly found on X [9]. IndoBERTweet has been utilized in several studies and has performed better than traditional models.. For example, study [21] conducted sentiment analysis on Indonesian TikTok reviews and found that IndoBERTweet outperformed an LSTM model, achieving an accuracy of 80% compared to 78%. Similarly, study [22] compared IndoBERTweet with a Support Vector Machine (SVM) for sentiment analysis related to the racing circuit construction in Indonesia, using data collected from X, reporting that IndoBERTweet achieved an F1-score of 88.4%, surpassing SVM's 82%.

While IndoBERTweet has demonstrated superior performance in various single-task sentiment analysis applications, there remains a limited research effort exploring its integration into multi-task learning frameworks for aspect-based sentiment analysis, particularly on informal and noisy Indonesian social media data. This gap motivates our study, which aims to combine IndoBERTweet's contextual embeddings with a lightweight CNN-based architecture in a multi-task setting to jointly predict aspects and sentiments in tweets related to the Mandalika International Circuit.[19]

III. METHODOLOGY

This study involves several stages, as shown in Figure 1.

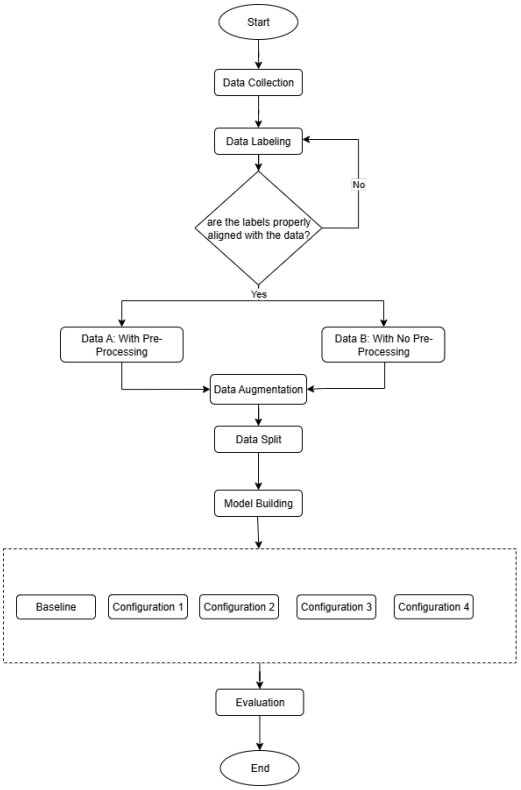

Fig. 1. Research Flow

Figure 1 illustrates the overall research workflow. The process begins with data collection, followed by data labeling for aspect and sentiment. A label verification step is then performed to check whether the labels align with the input data. If misalignment is found, the labeling process is repeated to ensure consistency. The workflow proceeds with data preprocessing and augmentation, followed by the splitting of the data into training, validation, and test sets. The model is then trained using a baseline and several configurations. Finally, model performance is evaluated. The following sections provide more detailed explanations of each step.

A. Data Collection

We collected 23,722 Indonesian tweets from X related to the Mandalika International Circuit from 2022 to 2024. The data crawling process used `tweet-harvest`, a Playwright-based tool that collected tweets using keywords and date filters. The keywords "MotoGP Mandalika" and "Sirkuit Mandalika" were used to retrieve the related tweets.

B. Data Annotation

At this stage, we perform human annotation to label the tweets with aspect and sentiment labels. We asked two university students, both native Indonesian speakers, to act as the annotators. Due to resource limitations, manual annotation was conducted on a randomly selected subset of 3,556 tweets from the original dataset. The tweets were divided between two annotators, with an intentional overlap of 250 tweets assigned to both. This overlapping set was used to assess inter-annotator agreement.

We perform the prescriptive paradigm data annotation. This paradigm aims to minimize subjectivity among annotators by defining rules outlined in the annotation guidelines [23]. We followed the sentiment analysis annotation guidelines developed by [24], including three sentiment labels: positive, negative, and neutral. The aspect guidelines were developed independently with four labels: "Ekonomi" (Economy), "Politik" (Politics), "Fasilitas" (Facility), and "Lainnya" (Others), as shown in Table I.

TABLE I. ASPECT ANNOTATION GUIDELINES

Aspect	Definition	Criteria
Economy	Matters related to impact, money circulation, business, national/regional revenue, or community activities.	Mentions UMKM (micro, small, and medium enterprises), tourism, sales, and investment.
		Discusses financial losses, gains, or impacts.
		Refers to development plans or economic branding of the area.
		Mentions regional revenue, foreign exchange, national/state budgets (APBN/APBD)
Politics	Matters related to political actors, government decisions, or national/local political dynamics concerning the circuit/event concerning the circuit/event	Sentences containing praise, criticism, sarcasm, or support towards government or political figures.
		Contains elements of policy, political promises, or comparisons between politicians/policies
		Contains elements of policy, political promises, or comparisons between politicians/policies
Facilitiy	Matters related to infrastructure, services, and physical comfort around the Mandalika MotoGP event.	Mentions circuit asphalt, stands, shuttle services, toiles, and road access.
		Discussion about repairs, construction, or facility readiness.
		Complaints or praises regarding cleanliness, traffic jams, transportation, and parking.
Others	Matters not specific to the other three aspects, such as general comments, jokes, admiration, or neutral news	A general or entertaining comment.
		Description of events, fan narratives, promotions, or light jokes.

Table I presents the aspect guidelines used for annotating the dataset. Each aspect was defined, with details stated in the corresponding criteria column. We maintained an iterative annotation process, refining guidelines based on annotator questions and disagreements, ensuring clarity and consistent labeling.

To evaluate the consistency of these annotations, Cohen's Kappa was calculated between the two annotators. The resulting kappa score of 0.90 indicates a strong level of agreement between the annotators [25]. Examples of the annotated data are presented in Table II.

TABLE II. RESULT OF ANNOTATED DATA

No.	Text	Aspect	Sentiment
1	Aspal Sirkuit Mandalika mengelupas dan rusak bikin pembalap luka-luka.	Fasilitas	Negative
2	Apa yang harus dipikirkan kami sebagai warga NTB sangat berterima kasih atas Sirkuit Mandalika itu membuktikan Pak Jokowi sangat peduli NTB.	Politik	Positive
3	Ada gak ya lomba balapan lain? Skala internasional selain MotoGP buat pemasukan pengelola Mandalika.	Ekonomi	Neutral
4	Akhir kata semangat buat para pembalap ditunggu untuk balapan selanjutnya.	Lainnya	Positive

Based on Table II, the first tweet, "Aspal Sirkuit Mandalika mengelupas dan rusak bikin pembalap luka-luka" (The Mandalika Circuit's asphalt is peeling and damaging, causing injuries to the riders), is labeled as Fasilitas-Negative. This sentence focuses on the facility of the Mandalika Circuit, specifically its asphalt. The sentiment expressed is negative due to the damage to the asphalt, which has caused injuries to the riders.

The second tweet, "Apa yang harus dipikirkan kami sebagai warga NTB sangat berterima kasih atas Sirkuit Mandalika itu membuktikan Pak Jokowi sangat peduli NTB" (What we, as people of NTB, should think is that we are very grateful for the Mandalika Circuit as it proves that Mr. Jokowi cares deeply about NTB), is labeled as Politik-Positive. This sentence expresses gratitude, a positive sentiment towards the 7th President of Indonesia, Mr. Joko Widodo, for his contribution to the development of the Mandalika Circuit in NTB.

The third tweet, "Ada gak ya lomba balapan lain? Skala internasional selain MotoGP buat pemasukan pengelola Mandalika" (Are there any other international-scale racing competitions besides MotoGP that could generate income for the Mandalika management?), is labeled as Ekonomi-Neutral because it explicitly discusses the potential for earning through other international-scale events without showing any strong emotion or judgment, instead simply posing a question. Therefore, this sentence is labeled as neutral.

The fourth tweet, "Akhir kata semangat buat para pembalap ditunggu untuk balapan selanjutnya" (Last but not least, good luck to the riders, looking forward to the next race), is labeled as Lainnya-Positive. This sentence conveys a positive sentiment by encouraging the riders, who are the implied audience. However, since no specific aspect is mentioned for the riders, it is labeled under "Lainnya".

C. Data Preprocessing

Data preprocessing is done to clean and turn raw data into an understandable format [26]. The preprocessing steps are as follows:

C.1 Case Folding and Special Character Removal

This process employs case folding to convert all letters in the dataset to lowercase, ensuring that words with different capitalizations are treated equally [26]. This process also involves removing all special characters, including links, usernames (@usernames), hashtags (#), extra spaces, and punctuation. This study uses regular expressions to clean the data. Table III shows the result of the basic text cleaning process.

TABLE III. RESULT OF CASE FOLDING AND CHARACTER REMOVAL

Original Text	Result
Wow, Sirkuit Mandalika bener-bener keren, gokil bgt deh, knp ga dpt tiketnya ya? wkwkwk :v	Wow sirkuit mandalika benerbener keren gokil bgt deh knp ga dpt tiketnya ya wkwkwk v

As shown in Table III, case folding was applied by changing the phrases "Wow," "Sirkuit," and "Mandalika" to lowercase. Then, punctuation removal was carried out by eliminating the characters "," "-" "?", and ":".

C.2 Slang Normalization

Slang normalization addresses the presence of slang, often considered noise in text processing [27]. During this process, we normalize using "Kamus Alay", a colloquial Indonesian lexicon developed by [28]. The slang dictionary consists of two columns, slang and formal. The corresponding formal form will replace any slang words found within the sentence.

Although the "Kamus Alay" was published seven years ago, it remains widely used in recent research for the normalization of slang in Indonesian social media texts [29], [30], [31]. Moreover, for newer slang that the dictionary may not cover, IndoBERTweet helps manage it through its contextual embeddings, which are trained on recent X data.

TABLE IV. RESULT OF SLANG NORMALIZATION

Original Text	Result
Wow, Sirkuit Mandalika bener-bener keren, gokil bgt deh, knp ga dpt tiketnya ya wkwkwk v	wow sirkuit mandalika benerbener keren gila banget deh kenapa enggak dapat tiketnya ya wkwkwk v

Table IV presents the results of slang normalization, which involves converting informal phrases to their formal

counterparts. For instance, “bgt” was changed to “banget” (so), “knp” was replaced with “kenapa” (why), and “ga” was changed to”enggak” (can not).

C.3 Short Word Removal

Short word removal prevents further noise by removing all words less than three characters [32].

TABLE V. RESULT OF SHORT WORD REMOVAL

Original Text	Result
wow sirkuit mandalika benerbener keren gila banget deh kenapa enggak dapat tiketnya ya wkwkwk v	wow sirkuit mandalika benerbener keren gila banget deh kenapa enggak dapat tiketnya wkwkwk

As presented in Table V, the short word removal was applied by deleting the phrase “v” that only has one character.

D. Data Augmentation

Data augmentation is a method for increasing the amount of data by adding modified copies of the existing data [33]. Our study purposely used this approach to address class imbalance, which was identified using Google for Developers' guidelines [34]. We identified the class distributions at two levels: first, at the general task level for sentiment and aspect separately, and second, at the finer granularity of each combined sentiment-aspect class, as shown in Table VI.

TABLE VI. ASPECT-SENTIMENT DISTRIBUTION BEFORE AUGMENTATION

Aspect	Sentiment (%)			Total Per Aspect (%)
	Positive	*Negative*	*Neutral*	
Ekonomi	65.98	24.57	9.45	17.51
Politik	33.51	45.90	20.95	11.84
Fasilitas	45.83	14.93	39.24	35.96
Lainnya	43.49	17.54	38.98	34.69
Total Per Sentiment (%)	46.51	22.02	31.47	

According to Table VI, the negative sentiment class has the smallest overall proportion, comprising just 22.02% of the sentiment categories. The “Ekonomi” and “Politik” aspects also show lower overall representation compared to “Fasilitas” and “Lainnya.” Within the “Ekonomi” aspect, neutral sentiment is particularly low at 9.45%, and the negative class also needs to be increased to balance the distribution. For the “Politik” aspect, all sentiment classes are increased due to lower counts and notable gaps compared to other aspects, even though positive sentiment exceeds 20%. In the “Fasilitas” and “Lainnya” aspects, only the negative sentiment classes are increased, as they hold the smallest shares within those categories. The augmentation process was carried out using two methods, as described in Sections D.1 and D.2.

D.1 Back-to-Back Translation

Back-to-back translation is a method that involves translating the original text into another language, and then translating it back to the original language. This process aims to rephrase the entire sentence in a new form [33].

To perform back-to-back translation, we use the neural machine translation called OPUS-MT [35]. Before using OPUS-MT on the data, we tested it using dummy data, which demonstrated its ability to handle slang and informal language effectively. During this process, we set the source language to “ID” (Indonesian) and the target language to “EN” (English).

TABLE VII. RESULT OF BACK-TO-BACK TRANSLATION

Source Language	Target Language	Result
Pagi ini niatnya mau beli tiket, eh udah habis aja, sedih.	This morning he was going to buy a ticket, uh just run out, sad.	Pagi ini dia akan membeli tiket, eh hanya kehabisan, sedih.

Table VII shows that the Source Language column contains raw text in the Indonesian language. This text was translated into English as the target language. Subsequently, the English translation was back-translated into Indonesian. The result obtained from this process is slightly more formal than the original text. However, it still preserves the original context.

D.2 Synonym Replacement

Synonym replacement is a method of replacing words in the original text with their synonyms [33]. We used the Indonesian synonym dictionary to perform this process. We excluded stopwords by utilizing a stopword dictionary to ensure the augmented results remain natural.

TABLE VIII. RESULT OF SYNONYM REPLACEMENT

Original Text	Result
Pemandangan di sekitar are Sirkuit Mandalika sangat indah untuk dilihat	Pemandangan di seputar distrik Sirkuit Mandalika amat artistik untuk dilihat

Table VIII shows the replacement of some phrases within the original text. The phrase “sangat” (very) was replaced with “amat”, and the phrase “indah” (beautiful) was replaced with “artistik”.

Following the augmentation processes, the dataset comprises 6,480 instances, with the final distribution shown in Table VIII.

TABLE IX. ASPECT-SENTIMENT DISTRIBUTION AFTER AUGMENTATION

Aspect	Sentiment (%)			Total Per Aspect (%)
	Positive	*Negative*	*Neutral*	
Ekonomi	48.91	29.42	21.67	23.29
Politik	40.75	35.19	24.07	21.93
Fasilitas	48.11	20.15	31.74	30.63
Lainnya	39.94	31.95	28.12	24.15
Total Per Sentiment (%)	44.71	28.46	26.84	

Table IX shows the distribution of sentiments across aspects after data augmentation. The overall sentiment balance has improved, with positive sentiment at 44.71%, negative at 28.46%, and neutral at 26.84%. Within aspects, distributions are more balanced compared to the original data. For instance, the “Politik” aspect still has a relatively high negative sentiment (35.18%), while “Fasilitas” now

has a stronger neutral component (31.74%). These results indicate that augmentation can address previously underrepresented classes.

E. Data Split

Data split is a process of splitting the data into a training and a test set [36]. In this process, the data is split into a training, validation, and test set with a ratio of 80%:10 %:10% respectively [37]. Table X shows the total for each set.

TABLE X. DATA SPLIT DISTRIBUTION

Train Set	Validation Set	Test Set
3,571	446	447

Based on the data split described in Table X, 80% of the data, or 3,571 samples, is used for training, 10% (446 samples) is reserved for validation, and the remaining 10% (447 samples) is used for testing.

F. Model Architecture

The model architecture comprises a multi-task learning framework that combines IndoBERTweet embedding with CNN. The model is designed to jointly predict aspect categories and sentiment polarities for each tweet, thereby enabling it to better understand the connection between the two tasks.

F.1 IndoBERTweet Embeddings

The model utilizes IndoBERTweet to create word embeddings that are explicitly trained on Indonesian data. In this study, the model employed "`indobertweet-base-uncased`". This enables the model to comprehend informal language, slang, and abbreviations commonly used on social media. These embeddings provide a foundational representation of the text, helping the model to understand the meaning and context of the words.

F.2 Convolutional Neural Network

Before finalizing the architecture, we experimented with different CNN configurations, varying the number of convolutional layers, kernel sizes, and feature dimensions. Based on the performance observed during testing, the current design delivered the best result. Therefore, this architecture was selected for the experiments on the architecture composition.

The embeddings are passed through CNN layers to extract local patterns, such as n-grams and key phrase patterns. Using multiple 1D convolutional layers with varying kernel sizes (2, 3, and 4) enables the model to capture different lengths of n-grams, which is crucial for understanding aspect-specific sentiments in short texts, such as tweets. The output of the convolutional layers is pooled using AdaptiveMaxPool1d, which reduces the sequence dimensions while preserving the most essential features.

Following the convolutional layers, the output is flattened and passed through a fully connected layer, which reduces the feature space to 256 dimensions. This shared layer helps combine the information from the convolutional layers, and dropout is applied to prevent overfitting.

The model employs a multi-task learning approach by predicting two tasks simultaneously: aspect and sentiment classification. The aspect classifier predicts the category of the aspect, while the sentiment classifier predicts the polarity of sentiment. Both tasks share the same lower layers, which helps the model learn common features and improve efficiency. The output for each task uses a softmax activation function to produce probability distributions for the aspect categories and sentiment polarities.

G. Experimental Setup

For the experiment, we performed two scenarios for the dataset as shown in Table XI.

TABLE XI. DATA SCENARIOS

Scenario	Description
1	No pre-processing and performing augmentation.
2	With pre-processing and performing augmentation.

Based on Table XI, the data is generally divided into two subsets: one with applied preprocessing and the other without. The subset without preprocessing aims to evaluate IndoBERTweet's performance directly on raw X data without modification or cleaning.

This study compares two approaches: (1) a baseline model with hyperparameter configuration adapted from a prior study, and (2) a model with hyperparameter tuning. Both models were trained on the same dataset and evaluated under the same conditions with different configurations. Each model configuration was run five times, and both the average and standard deviation of the F1-scores, along with their corresponding p-values, were calculated.

G.1 Baseline Model

We adopt the hyperparameter used in [38] for the baseline model with a batch size of 32, a learning rate of 2e-5, and a dropout rate of 0.4. In their study, the model achieved 97% accuracy with a 95% F1-score, which serves as a reference point for evaluating the performance of our proposed approach. This provides a benchmark to compare the effectiveness of the proposed multi-task model.

G.2.1 Hyperparameter Tuning

Building on the baseline setup, we further explored several hyperparameter configurations manually to optimize the performance of the proposed multi-task ABSA model. The configurations used in the experiments are shown in Table XII.

TABLE XII. HYPERPARAMETER TUNING CONFIGURATION

Configuration	Learning Rate	Batch	Dropout
1	1e-5	16	0.3
2	2e-5	16	0.5
3	2e-5	16	0.3
4	2e-5	32	0.3

As shown in Table XII, the hyperparameters are tuned with different learning rates (1e-5 and 2e-5), batch sizes (16 and 32), and dropout (0.3 and 0.5).

G.2.2 Influence of Data Augmentation

To assess the influence of data augmentation, we first conducted all experiments using augmented data. We then re-tested the best-performing configuration without augmentation to enable a clear comparison of its impact on model performance.

H. Evaluation

H.1 Precision, Recall, Fscore

Three primary evaluation metrics assess the model's performance: precision, recall, and F1-score. These metrics are calculated based on the number of correct and incorrect predictions made by the model. In classification tasks, the prediction results can be categorized as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP refers to the number of instances correctly predicted as positive, FP is the number of cases wrongly predicted as positive, TN represents the number of cases correctly predicted as negative, and FN is the number of cases incorrectly predicted as negative [26]. By using these, the precision, recall, and F1-score are defined as Equations 1, 2, and 3, respectively [39].

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Fscore = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

According to Equation 1, precision represents the proportion of correctly identified positive predictions among all instances predicted as positive. A high precision indicates that the model makes few false positive errors. As described in Equation 2, recall measures the proportion of actual positive cases that are correctly predicted by the model, reflecting its ability to detect relevant instances. Equation 3 defines the F1-score as the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. A high F1-score suggests strong performance in both aspects.

Since this study involves multiple classes with varying levels of representation, the weighted F1-score was used as the primary evaluation metric. This metric calculates the F1 Scores for each class and then takes the average, weighted by the number of actual instances (support) in each class. Thus, it reflects the model's overall performance.

H.2 Mean, Standard Deviation, and T-Pair Test

To assess the consistency of each configuration's performance, this study reports the mean and standard deviation (SD) of the F1-scores across five runs. The mean provides a measure of central tendency, representing the average performance of a configuration. At the same time, the standard deviation quantifies the variability of the scores, indicating how stable or spread out the results are from the mean. A lower standard deviation suggests more consistent performance.

To further assess whether there is a statistically significant difference in performance between the two model configurations, a paired sample t-test (also known as a paired t-test) was performed. This test compares the average F1-scores from the same configuration run under two different conditions, Scenario 1 and Scenario 2. Since the scores come from matching runs, the test looks at whether the differences between them are consistent enough to be considered statistically significant. If the p-value is below 0.05, it means the difference is likely not due to chance. On the other hand, a p-value above 0.05 suggests that there is no substantial evidence to indicate that one scenario performs better than the other on average [40].

I. Configuration Performance Summary

For further discussion in this study, we compare all configurations and select the best one for detailed analysis. The selection is based on the mean and standard deviation of the F1-scores, obtained from five independent runs per configuration. The summarized results are presented in Table XIII.

TABLE XIII. MODEL PERFORMANCE RESULT

Scenario 1		
Configurations	***Mean***	***Standard Deviation***
Baseline	57.67	0.137
Config 1	58.39	0.185
Config 2	58.98	0.021
Config 3	61.85	0.660
Config 4	57.84	0.202
Scenario 2		
Configurations	***Mean***	***Standard Deviation***
Baseline	55.93	0.152
Config 1	58.34	0.152
Config 2	58.72	0.125
Config 3	58.37	0.120
Config 4	58.48	0.062

Based on the performance comparison shown in Table XIII, Configuration 2 was selected for further analysis. While Configuration 3 in Scenario 1 achieved the highest mean F1-score (61.85), it also exhibited the largest standard deviation (0.660), indicating that its performance varied significantly across different runs. In contrast, Configuration 2 consistently delivered high mean F1-scores in both Scenario 1 (59.98) and Scenario 2 (58.72), with relatively low standard deviations of 0.021 and 0.125. This balance of strong performance and stability makes Config 2 the most reliable configuration for subsequent evaluations.

IV. RESULTS AND DISCUSSIONS

A. Model Performance

To further analyze model performance, we conducted a paired t-test to compare F1-scores for each aspect-sentiment class across the two experimental scenarios. This comparison aims to determine whether the change

introduced in Scenario 2 resulted in statistically significant improvements or declines in performance compared to Scenario 1. Table XIV presents the average F1-scores for each class in both scenarios, along with the p-values resulting from the paired t-test.

TABLE XIV. SCENARIO COMPARISON OF CONFIG 2: MEAN F1-SCORES AND PAIRED T-TEST RESULTS

Aspect-Sentiment Class	Scenario 1	Scenario 2	t-pair p-value
Ekonomi-Positive	65.32	71.31	0.016
Ekonomi-Negative	73.46	73.64	0.228
Ekonomi-Neutral	54.52	49.73	0.119
Politik-Positive	71.80	72.43	0.254
Politik-Negative	76.35	66.05	0.003
Politik-Neutral	65.92	55.25	0.009
Fasilitas-Positive	60.03	58.64	0.013
Fasilitas-Negative	60.42	67.32	0.008
Fasilitas-Neutral	55.41	57.01	0.435
Lainnya-Positive	44.81	45.71	0.249
Lainnya-Negative	59.27	52.82	0.310
Lainnya-Neutral	46.21	42.26	0.974

From Table XIV, several aspect-sentiment classes exhibit significant differences, notably "Ekonomi-Positive" ($p = 0.016$), "Politik-Neutral" ($p = 0.003$), "Politik-Neutral" ($p = 0.009$), "Fasilitas-Positive" ($p = 0.013$), and "Fasilitas-Negative" ($p = 0.008$), suggesting that the changes implemented in Scenario 2 had a measurable effect on these classes. In contrast, other classes such as "Ekonomi-Negative", "Politik-Positive", and "Lainnya-Neutral" show no significant change ($p>0.05$), indicating stable performance across scenarios. These findings support the notion that model improvements may be class-dependent and should be interpreted in the context of both statistical significance and overall classification consistency.

Furthermore, when comparing the overall F1-scores between the two scenarios using a paired t-test, the resulting p-value was 0.7178, indicating that, on average, there is no statistically significant difference in performance across the entire set of predictions.

B. Influence of Data Augmentation

Building on the results from Section A, we examined the influence of data augmentation by running an experiment without pre-processing using Configuration 2. The value presented in Table XV represent the mean results obtained from five separate runs.

TABLE XV. METRIC COMPARISON WITH AND WITHOUT AUGMENTATION

Without Augmentation		
Aspect-Sentiment	***Precision***	***Recall***
Ekonomi-Positive	67.52	66.32
Ekonomi-Negative	51.84	61.43
Ekonomi-Neutral	33.00	23.33
Politik-Positive	42.17	28.89
Politik-Negative	52.28	44.80
Politik-Neutral	44.46	54.54
Fasilitas-Positive	52.68	62.77
Fasilitas-Negative	49.54	57.27
Fasilitas-Neutral	55.10	56.07
Lainnya-Positive	50.99	40.91
Lainnya-Negative	52.84	46.67
Lainnya-Neutral	38.69	36.26
With Augmentation		
Aspect-Sentiment	***Precision***	***Recall***
Ekonomi-Positive	74.74	58.21
Ekonomi-Negative	66.09	83.33
Ekonomi-Neutral	44.87	70.00
Politik-Positive	78.29	66.43
Politik-Negative	84.48	69.68
Politik-Neutral	59.01	75.38
Fasilitas-Positive	57.84	62.55
Fasilitas-Negative	48.01	81.54
Fasilitas-Neutral	55.07	56.34
Lainnya-Positive	49.93	41.09
Lainnya-Negative	54.48	65.38
Lainnya-Neutral	55.20	41.71

Table XV compares the model's performance across aspect-sentiment classes with and without data augmentation, using precision and recall as evaluation metrics. In this analysis, scores below 50.0 are considered low. Overall, data augmentation appears to bring noticeable improvements, especially in recall across several classes.

For instance, the "Ekonomi-Negative" class shows a sharp increase in recall from 61.43 to 83.33, along with a moderate rise in precision from 51.84 to 66.09. This suggests the model becomes more sensitive in identifying negative economic sentiments after augmentation.

A similar pattern is seen in the "Fasilitas-Negative" class, where recall jumps from 57.27 to 81.54 while precision remains relatively stable, though still under 50.0. This indicates that while the model still struggles somewhat with precision, it's significantly better at capturing actual negative instances after augmentation. The "Politik-Negative" class also sees a substantial boost in both metrics, with precision improving from 52.28 to 84.48 and recall increasing from 44.80 to 69.68, indicating enhanced consistency and accuracy.

Meanwhile, the "Ekonomi-Neutral" class, which initially had very low scores (precision = 33.00, recall = 23.33), improves significantly to 44.87 and 70.00, respectively. Although precision is still below the moderate range, the increase in recall indicates a significant reduction in false negatives. "Lainnya-Neutral" also improves slightly in both precision (from 38.69 to 55.20) and recall (from 36.26 to 41.71), reflecting a modest but positive impact.

Other classes like "Lainnya-Positive" and "Politik-Positive" also show gains, particularly "Politik-Positive", which increases dramatically from 42.17/28.89 to 78.29/66.43 in precision and recall.

C. Qualitative Analysis

Based on the evaluation in Section A, Scenario 1, where the model used raw, unprocessed text, performed better than Scenario 2, which involved standard preprocessing. This outcome demonstrates that IndoBERTweet is effective with informal language, as it was pre-trained on tweet data that commonly includes slang, repeated letters, and non-standard spelling. Cleaning the text too much may remove proper signals that help the model understand

sentiment or intention. Furthermore, this is supported by the t-test conducted on the final F1-scores between the two scenarios, which yielded a p-value of 0.7178, indicating that, on average, there is no statistically significant difference in overall performance across the two scenarios. While some aspect-sentiment classes showed meaningful differences individually, the aggregate result suggests that aggressive preprocessing does not lead to consistent improvement and may even reduce the model's ability to generalize across informal expressions found in social media text.

Additionally, the CNN layer in the model helps identify important local patterns, such as short phrases or repeated expressions, often found in tweets. When the text is cleaned too aggressively, those patterns may be lost, reducing the model's ability to capture relevant features. Since this model also employs a multi-task approach, predicting both aspects and sentiment simultaneously, removing useful text information can impact both tasks simultaneously.

The results in Section B also demonstrate that data augmentation had a positive impact, particularly in enhancing the model's ability to handle underrepresented classes. By introducing variation, augmentation enables the model to learn a broader range of expressions and sentence structures, thereby making it more robust.

Even so, some classes remain difficult. The "Ekonomi-Neutral" and "Fasilitas-Negatice" class, for example, often shows high recall but low precision. This suggests that while the model can find many correct instances, it also mislabels others as belonging to this class. The "Lainnya" aspect also consistently shows low performance, possibly due to unclear boundaries or a lack of training data.

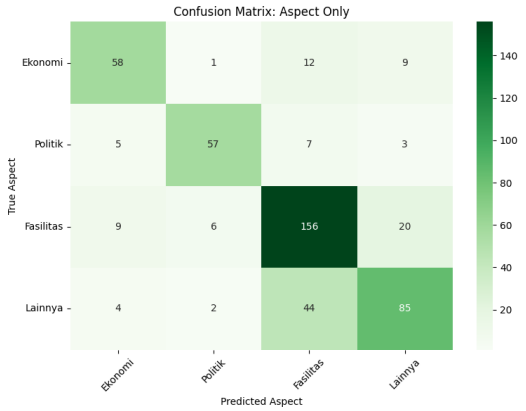

Fig. 2. Confusion Matrix of Sentiment in Scenario 1

The confusion matrix in Figure 2 shows the difficulty in classifying the "Lainnya" aspect. Of all aspect labels, "Lainnya" has the highest number of misclassifications, with 20 instances incorrectly predicted as Fasilitas, and additional confusion with "Ekonomi" and "Politik"

Unlike other aspects with more explicit semantic cues, "Lainnya" is a broad category that introduces ambiguity during training. The lack of specific keywords or consistent patterns makes it more challenging for the model to accurately identify and classify this aspect. This confusion is particularly evident with "Fasilitas", suggesting either semantic overlap between these categories or the need for more explicit aspect definitions. This often results in low recall, where the model misses many actual instances, and low precision, where unrelated inputs are wrongly predicted as "Lainnya".

In addition, for several aspects, the Positive and Neutral sentiment classes consistently record the two lowest F1-scores within the same aspect group. For example, in the "Ekonomi" aspect, the Positive and Neutral classes have the lowest scores compared to their Negative class. A similar trend is observed in other aspects, such as "Lainnya" and "Fasilitas".

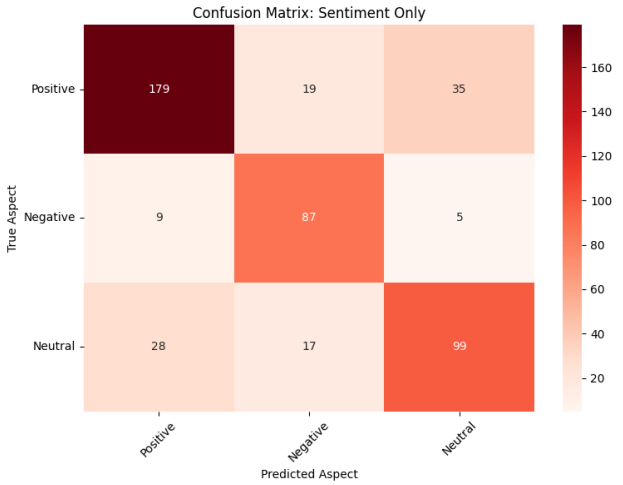

Fig. 3. Confusion Matrix of Sentiment in Scenario 1

Figure 3 shows that the model misclassifies 28 Positive cases as Neutral and 35 Neutral cases as Positive. This confusion suggests that the boundary between the two sentiments is often unclear in the dataset. Informal language on social media frequently includes implicit or understood positive cues, which can obscure the boundary between Neutral and Positive sentiment.

As illustrated in Table XVI, further insights were gained by examining the model's output.

TABLE XVI. MISCLASSIFICATION EXAMPLE

Tweet	Actual Label	Predicted Label
Belum Pernah Finis Setiap Balapan di Sirkuit Mandalika Marc Marquez Ungkap Misi Besar Tahun Ini #SukseskanGPMandalika Jamin Kelancaran Agenda	Lainnya-Neutral	Fasilitas-Positive

Table XVI shows the text "Belum Pernah Finis Setiap Balapan di Sirkuit Mandalika Marc Marquez Ungkap Misi Besar Tahun Ini #SukseskanGPMandalika Jamin Kelancaran Agenda" (Never Finished Any Race at the Mandalika Circuit, Marc Marquez Reveals a Big Mission This Year #SupportGPMandalika Ensuring the Smoothness of the Agenda) is labeled as "Lainnya-Neutral". The actual label focuses on Marc Marquez as its central aspect, with an overall tone that is factual rather than

emotional. However, the model misclassified it as "Fasilitas-Positive," likely due to its emphasis on the phrase "Jaminan Kelancaran Agenda" and the use of "Sukseskan," both of which imply a positive outcome and relate to organization or infrastructure.

This misclassification may occur when multiple ideas are expressed in a single tweet, resulting in class overlap. This overlap makes it difficult for the model to assign a single, clear aspect-sentiment label, especially when different parts of the tweet convey different tones or focus on other topics.

D. Comparison with Previous Work

Our proposed approach was subsequently tested on the same dataset used in a prior study [10]. The dataset used comprises two aspects, "Ekonomi" (Economy) and "Politik" (Politics), along with three sentiment classes, which are similar to those in our study. This experiment followed Scenario I (no preprocessing with an augmentation) and utilized the best-performing configuration, Configuration 2. The results are shown in Table XVII.

TABLE XVII. PROPOSED APPROACH COMPARISON

Approach	Accuracy	
	Aspect	***Sentiment***
CNN+FastText Embeddings (Manuaba et.al.)	84%	72%
CNN + IndoBERTweet Embeddings (Our Approach)	88%	83%

Based on Table XVII, the CNN model with FastText embeddings achieved accuracies of 84% for aspect classification and 72% for sentiment classification, as stated in [10]. By using our proposed approach, the model achieved a higher accuracy of 88% for aspect classification and 83% for sentiment classification.

The higher performance achieved using the proposed approach may be due to the use of contextual embeddings, such as IndoBERTweet, which can better understand the meaning and variations in informal social media language compared to static embeddings like FastText used in the prior study.

V. CONCLUSIONS

This study presents an ABSA approach to Indonesian-language tweets related to the Mandalika International Circuit. We proposed a multi-task learning framework combining a CNN with IndoBERTweet embeddings. The model was evaluated under two experimental scenarios: one using raw, unprocessed data, with augmentation, and the other incorporating text preprocessing and data augmentation. Results indicate that Scenario 1 outperformed Scenario 2 in multiple classes, suggesting that IndoBERTweet performs better with informal and noisy text, which reflects the nature of its pretraining data from X/Twitter. Excessive text cleaning may strip valuable cues such as slang, repetition, and casual phrasing that contribute to the accurate interpretation of sentiment and aspect. Among all configuration tests, Configuration 2 (learning rate = 2e-5, batch size = 16, and dropout = 0.5) achieved the best overall performance with a high F1-score and low standard deviation, demonstrating both strong accuracy and stability.

Statistical testing using a paired t-test revealed no significant difference in the overall F1-scores between the two scenarios ($p = 0.7178$), reinforcing the observation that preprocessing does not consistently provide performance gains. Additionally, class-level analysis revealed statistically significant changes in certain classes, such as "Ekonomi-Positive", "Politik-Neutral", and "Fasilitas-Negative". Others, like "Ekonomi-Negative" and "Lainnya-Neutral", remained unaffected. This highlights that model improvements are class-dependent and should be interpreted within the context of both statistical significance and classification consistency.

Furthermore, experiments comparing performance with and without data augmentation showed clear improvement, particularly in recall, across underrepresented classes. This indicates that augmentation techniques helped the model generalize better and reduced bias toward dominant classes by exposing the model to more varied linguistic patterns. However, even with augmentation, certain classes, such as "Lainnya", "Ekonomi-Neutral", and "Fasilitas-Negative", continued to exhibit low precision or recall, indicating persistent challenges in distinguishing these categories. This is likely caused by an imbalanced class distribution, overly broad aspect labels that lack precise definitions, and the presence of multiple aspects or sentiments in a single tweet.

Additionally, when evaluated on the dataset from a prior study, the proposed approach achieved higher weighted F1-scores than the preceding method, with 88% for aspect and 83% for sentiment classification. This may be due to the use of contextual embeddings, such as IndoBERTweet, which better handle informal social media language than static embeddings.

This study contributes to the knowledge of utilizing a multi-task learning framework with CNN and IndoBERTweet for aspect-based sentiment analysis, particularly on tweet data related to the Mandalika International Circuit. Additionally, this work has created a new dataset for aspect-based sentiment analysis, comprising 23,722 unannotated and 3,556 annotated datasets that can be utilized by other researchers for future studies.

Nevertheless, the study has limitations, particularly in the annotation process. Involving domain experts in future labeling efforts may enhance annotation consistency and semantic clarity. For future work, we recommend refining broad aspect categories such as "Lainnya" into more granular sub-aspects to improve classification performance. Moreover, to address the issue of multiple sentiments or aspects within a single input, future research may explore sentence chunking strategies or adopt multi-label classification frameworks.

VI. REFERENCES

[1] I. Zitri, C. Kurniawan, and T. Octastefani, "Public-Private Partnerships: In the Development of the Mandalika Circuit, Indonesia," *Journal of Governance and Public Policy*, vol. 11, no. 2, pp. 156–166, Jun. 2024, doi: 10.18196/jgpp.v11i2.17714.

[2] Z. Rahmadi, Hariyadi, and A. Pracoyo, "Economic Feasibility Analysis of Investment in International Circuit Development in the Mandalika Special Economic Zone (SEZ)," *RESEARCH REVIEW International Journal of Multidisciplinary*, vol. 9, no. 6, pp. 143–152, Jun. 2024, doi: 10.31305/rrijm.2024.v09.n06.020.

[3] M. Noviana, E. Malihah, and S. Komariah, "The Impact of Mandalika Circuit Development On Socio-Cultural Changes : A Systematic Literature Review," *Jupiis: Jurnal Pendidikan Ilmu-Ilmu SosiaL*, vol. 15, no. 1, p. 32, Jun. 2023, doi: 10.24114/jupiis.v15i1.42298.

[4] E. Cano-Marin, M. Mora-Cantallops, and S. Sánchez-Alonso, "Twitter as a predictive system: A systematic literature review," *J Bus Res*, vol. 157, Mar. 2023, doi: 10.1016/j.jbusres.2022.113561.

[5] M. Woodward, "Twitter User Statistics 2025: What Happened After 'X' Rebranding?," Search Logistic.

[6] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01054

[7] Y. Zhang and Q. Yang, "A Survey on Multi-Task Learning," Dec. 01, 2022, *IEEE Computer Society*. doi: 10.1109/TKDE.2021.3070203.

[8] Wang. Wei and J. Gang, "Application of Convolutional Neural Network in Natural Language Processing," 2018, doi: 10.1109/ICISCAE.2018.8666928.

[9] F. Koto Jey Han Lau Timothy Baldwin, "INDOBERTWEET: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization." [Online]. Available: https://huggingface.co/huseinzol05/

[10] I. B. Ryand, W. Manuaba, R. Dwiyansaputra, and M. Z. Hamidi, "Pendekatan Sentimen Berbasis Aspek Pada Ulasan Sirkuit Mandalika Menggunakan Cnn dan Representasi Fasttext (Aspect-Based Sentiment Approach to Mandalika Circuit Reviews Using CNN and FastText Representation)," Mar. 2025. [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/

[11] P. R. Amalia and E. Winarko, "Aspect-Based Sentiment Analysis on Indonesian Restaurant Review Using a Combination of Convolutional Neural Network and Contextualized Word Embedding," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, p. 285, Jul. 2021, doi: 10.22146/ijccs.67306.

[12] D. I. Anggraeni, P. D. Rizki, M. B. Setiawan, and A. B. Handayani, "Aspect-Based Sentiment Analysis for Indonesian Tourist Attraction Reviews Using Bidirectional Long Short-Term Memory," 2023.

[13] S. Cahyaningtyas, D. Hatta Fudholi, and A. Fathan Hidayatullah, "Deep Learning for Aspect-Based Sentiment Analysis on Indonesian Hotels Reviews," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Aug. 2021, doi: 10.22219/kinetik.v6i3.1300.

[14] H. T. Ismet, T. Mustaqim, and D. Purwitasari, "Aspect Based Sentiment Analysis of Product Review Using Memory Network," *Scientific Journal of Informatics*, vol. 9, no. 1, pp. 73–83, May 2022, doi: 10.15294/sji.v9i1.34094.

[15] F. Said and L. Parningotan Manik, "Aspect-Based Sentiment Analysis on Indonesian Presidential Election Using Deep Learning," *Paradigma*, vol. 24, no. 2, pp. 160–167, 2022, doi: 10.31294/p.v24i2.1415.

[16] H. Nguyen and K. Shirai, *A Joint Model of Term Extraction and Polarity Classification for Aspect-Based Sentiment Analysis*. IEEE, 2018.

[17] T. Uyen Tran, H. Thanh Ti Hoang, P. Hoai Dang, M. Riveill, M. Riveill Multitask Aspect, and H. Thanh Thi Hoang, "Multitask Aspect Based Sentiment Analysis with Integrated Bidirectional LSTM & CNN Model", doi: 10.1145/3440749.3442656ï.

[18] M. N. Hakim, S. A. I. Alfarozi, and P. I. Santosa, "Multi-Task Learning Aspect Based Sentiment Analysis with BERT," in *ICITEE 2024 - Proceedings of the 16th International Conference on Information Technology and Electrical Engineering 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 264–269. doi: 10.1109/ICITEE62483.2024.10808387.

[19] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative Study of CNN and RNN for Natural Language Processing," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.01923

[20] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What we know about how BERT works," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2002.12327

[21] J. C. Setiawan, K. M. Lhaksmana, and B. Bunyamin, "Sentiment Analysis of Indonesian TikTok Review Using LSTM and IndoBERTweet Algorithm," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 774–780, Aug. 2023, doi: 10.29100/jipi.v8i3.3911.

[22] H. M. Lee and Y. Sibaroni, "Comparison of IndoBERTweet and Support Vector Machine on

Sentiment Analysis of Racing Circuit Construction in Indonesia," *Jurnal Media Informatika Budidarma*, vol. 7, no. 1, p. 99, Jan. 2023, doi: 10.30865/mib.v7i1.5380.

[23] P. Röttger, B. Vidgen, D. Hovy, and J. B. Pierrehumbert, "Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.07475

[24] A. D. Latief, A. Jarin, M. T. Uliniansyah, E. Nurfadhilah, and D. I. N. Afra, "A Proven Sentiment Annotation Guideline for Indonesian Twitter Data," in *Proceedings - 2023 10th International Conference on Computer, Control, Informatics and its Applications: Exploring the Power of Data: Leveraging Information to Drive Digital Innovation, IC3INA 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 31–36. doi: 10.1109/IC3INA60834.2023.10285807.

[25] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem Med (Zagreb)*, pp. 276–282, 2012, doi: 10.11613/BM.2012.031.

[26] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 746–757, 2023, doi: 10.26555/jiteki.v9i3.26490.

[27] A. Maulana Barik, R. Mahendra, and M. Adriani, "Normalization of Indonesian-English Code-Mixed Twitter Data." [Online]. Available: https://blog.swiftkey.com/celebrating-international-

[28] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 226–229. doi: 10.1109/IALP.2018.8629151.

[29] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, "Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 534–539, 2022, doi: 10.14569/IJACSA.2022.0130665.

[30] F. F. Rachman, R. Nooraeni, and L. Yuliana, "Public Opinion of Transportation integrated (Jak Lingko), in DKI Jakarta, Indonesia," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 696–703. doi: 10.1016/j.procs.2021.01.057.

[31] M. Z. Rahman, Y. A. Sari, and N. Yudistira, "Analisis Sentimen Tweet COVID-19 menggunakan Word Embedding dan Metode Long Short-Term Memory (LSTM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 11, pp. 5120–5127, Nov. 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[32] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Applied Sciences (Switzerland)*, vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178765.

[33] B. Li, Y. Hou, and W. Che, "Data Augmentation Approaches in Natural Language Processing: A Survey," *AI Open*, vol. 3, pp. 71–90, Jan. 2022, doi: 10.1016/j.aiopen.2022.03.001.

[34] G. F. Developers, "Datasets: Imbalanced Datasets." Accessed: Nov. 10, 2024. [Online]. Available: https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets

[35] J. Tiedemann and S. Thottingal, "OPUS-MT-Building Open Translation Services for The World," 2020. [Online]. Available: http://opus.nlpl.eu

[36] E. Jain, J. Neeraja, B. Banerjee, and P. Ghosh, "A Diagnostic Approach to Assess the Quality of Data Splitting in Machine Learning," Jun. 2022, [Online]. Available: http://arxiv.org/abs/2206.11721

[37] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," 2024. [Online]. Available: www.ijacsa.thesai.org

[38] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences (Switzerland)*, vol. 15, no. 1, Jan. 2025, doi: 10.3390/app15010172.

[39] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The Impact of Features Extraction on The Sentiment Analysis," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 341–348. doi: 10.1016/j.procs.2019.05.008.

[40] A. I. Talika *et al.*, "On Paired Samples T-Test: Applications Examples and Limitations," *International Journal for Multidisciplinary Research (IGNATIAN)*, vol. 2, Apr. 2024.