Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

Multi-Aspect Sentiment Analysis on Gojek Application Reviews Using CNN-LSTM Method

Pujiaty Rezeki Saragih*, Yuliant Sibaroni, Sri Sulyani Prasetyowati

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: [1,*]pujiatysaragih@student.telkomuniversity.ac.id,[2]yuliant@telkomuniversity.ac.id,[3]srisuryani@telkomuniversity.ac.id
Correspondence Author Email: pujiatysaragih@student.telkomuniversity.ac.id
Submitted: **04/06/2025**; Accepted: **30/06/2025**; Published**: 30/06/2025**

Abstract−Since its initial release in 2010, Gojek has remained the most used online transport service by Indonesians. Multi-aspect sentiment analysis is a method applied to determine user sentiment towards various specific aspects in their comments. By applying this method, there will be deeper understanding of user views regarding various components of the Gojek service. The method employed was data scraping from web crawling of Google Play Store user reviews and data preprocessing, i.e., cleaning, case folding, tokenizing, stopword removal, normalization, and stemming. A hybrid CNN-LSTM model was employed since it is capable of extracting spatial features using CNN and long-term dependencies using LSTM. The seven most crucial aspects of the Gojek service, i.e., access, time, comfort, information, customer service, availability, and safety, were the central themes of this research. The main objective of this research is to analyze user sentiment across these key aspects using a deep learning-based multi-task approach, in order to gain actionable insights for improving service quality. The performance of the models was evaluated on accuracy as the primary metric, and the experiments attempted three model sizes: 32, 64, and 128 hidden units. Among them, the 64-unit model performed best overall consistently, with both aspect and sentiment classification accuracy being satisfactory. While the 128-unit model achieved slightly better accuracy on some sentiment tasks, it exhibited overfitting. The 64-unit model, however, gave the most balanced results and the best trade-off between model complexity and performance. The findings show the potential of multi-task deep learning approaches to extract valuable insights from user reviews. Such findings can be highly valuable to aid business strategy formulation and service quality improvement, and ultimately greater customer satisfaction, as well as consolidate Gojek's market dominance in Indonesia's online transport business.

Keywords: Sentiment Analysis; Multi-Aspect; CNN-LSTM; Gojek Reviews

1. INTRODUCTION

Gojek, established in 2010, is Indonesia's most used online transportation platform [1]. Being a pioneer in the country's ride-hailing industry [2], Gojek has a broad portfolio of services that cater to users' day-to-day activities ranging from transport, food delivery, logistics to digital payments. The company's repeated wins of service excellence awards by the Top Brand Awards for the past five years is a testament to its strength and commitment to service excellence [3].

Collection of user feedback, particularly from platforms such as Google Play, yields insightful information about experience and user opinion [4]. They form a robust corpus of unstructured information, which can be leveraged through sentiment analysis processes to determine the acceptance of people towards a service or product. Sentiment analysis enables wordy sentiments to be classified as positive, negative, or neutral [5] and provides business organizations with feedback that helps improve services and make better decisions [6].

Although sentiment analysis has been widely applied in the evaluation of digital services, a majority of previous studies have leaned towards overall sentiment classification. This prevents them from being capable of capturing users' nuanced opinions across different aspects of a service. In reality, users tend to have opinions towards multiple attributes—such as accessibility, information presence, customer support, and comfort—each of which can have a distinct impact on satisfaction. This points to the necessity of multi-aspect sentiment analysis (MASA) in achieving a richer and actionable understanding of user sentiments.

Several machine learning methods have been employed in previous sentiment analysis studies. Naïve Bayes, for example, was reported to deliver fairly low accuracy when processing Gojek user reviews, mainly because it is not well-equipped to handle the context complexity of long sentences [7]. While Support Vector Machine (SVM) was better—at a highest accuracy of 89.82% on Maxim service review sentiment analysis [4]—its computationally inefficacy with large data sets is a significant disadvantage. K-Nearest Neighbors (KNN) algorithm, which was optimized using Grid Search CV, had an accuracy of 83% on Gojek user reviews based on Twitter [8], but it also is not scalable since it is computationally intensive with increasing size of the dataset.

Deep learning techniques have shown improved performance in text data processing. CNN, for instance, when combined with feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and FastText, achieved 93.14% accuracy on content-related aspects in TikTok reviews [9]**.** However, CNN models are inherently restricted in their capacity to identify long-term semantic relationships in text. To address this, researchers have integrated Recurrent Neural Networks, i.e., Long Short-Term Memory (LSTM), which possess superior sequential data modeling and long dependency capabilities. An experiment with RNN-LSTM with BERT embeddings was found to have excellent performance, 95% accuracy for business features, 91% for content, and 85% for features [10]. LSTM models, however, typically require longer training time.

To fill the individual weakness of CNN and LSTM, researchers have attempted the combination of CNN-LSTM, leveraging the local feature extraction capability of CNN and the sequential modeling strength of LSTM. The hybrid proved to be beneficial for multi-aspect sentiment analysis. For example, CNN-LSTM on Bukalapak reviews

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

achieved 93.91% accuracy [11]. The hybrid approach is best for problems that must identify granular features and deep context, and as such, it is a perfect method to employ for analysis of advanced user reviews.

From this study, a CNN-LSTM hybrid model is proposed in attempting to conduct multi-aspect sentiment analysis on user reviews for the Gojek app gathered from the Google Play Store. To the best of our knowledge, no existing literature brought these techniques together in one framework specifically for Gojek, nor experimented on as many as seven different service factors simultaneously—i.e., access, time, availability, safety, information, customer service, and comfort. This approach is most likely to yield a deeper understanding of how users evaluate different dimensions of Gojek's services, which is crucial for both research study and business usage.

The novelty of this study does not just reside in its methodology, which integrates CNN for spatial feature learning and LSTM for learning long-term contextual dependencies [12], but also in its domain specificity. Previous research had centered on single-aspect sentiment classification, employed simpler models such as Naïve Bayes [7] or had been conducted on other services or platforms such as Maxim [4], Bukalapak [11], and TikTok [9]. The current study fills this gap by offering a multi-aspect, deep learning–driven sentiment analysis tailored specifically for Gojek, one of the biggest digital platforms in Indonesia, and by analyzing actual user-generated data from a high-volume, real-world dataset.

The aim of this study is to develop a more comprehensive sentiment analysis by making the most out of the CNN-LSTM hybrid model to recognize both local and distant textual features. The research will be able to enhance sentiment classification accuracy and offer practical insights into user satisfaction for specific service types. These insights can be applied to inform strategic business decisions, service quality improvements, and customer retention policies. Finally, this paper will help Gojek maintain its competitive advantage in the rapidly evolving Indonesian online transport industry.

2. RESEARCH METHODOLOGY

2.1. Research Stages

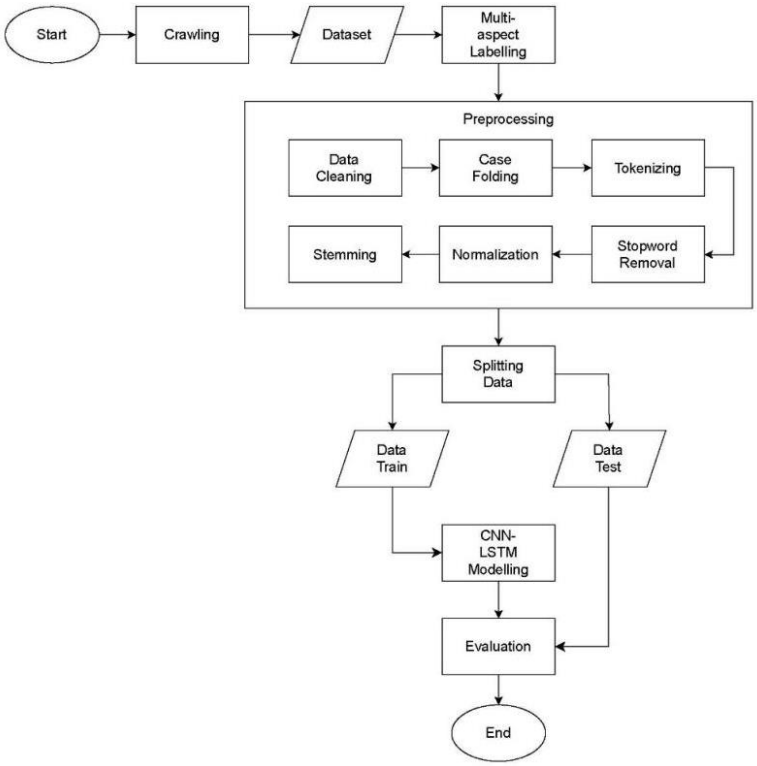

Figure 1. Research Flowchart

Sentiment analysis in this study is conducted using a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model to determine the sentiment of user reviews in different service fields of the Gojek app. The research process is directed by a straightforward and systematic sequence, which is graphically depicted in Figure 1. The steps begin with Data Crawling, which is the first step to collect user reviews from the Google Play Store. These reviews are then collected systematically into an organized dataset that is the foundation of all the following analysis.

Following data collection, Multi-Aspect Labeling is performed. This is where all the reviews are tagged with tags based on some aspects of the content of the Gojek service being reviewed. These aspects are significant in the analysis and comprise access, information, customer service, and comfort. By labeling these service aspects, this

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

research endeavors to determine what users think of these specific service facets, hence being able to harvest more detailed and actionable insights from the feedbacks. Multi-aspect tagging is required for determination of users' sentiment on a deeper level beyond the straightforward positive or negative sentiments.

The dataset then undergoes Preprocessing following the tagging process, where the text data is made ready for application in training the CNN-LSTM model. Preprocessing needs to be performed so that the data becomes more correct and in standard and readable form. Data Cleaning is then done initially, where irrelevances such as special characters, HTML tags, and unwanted white spaces are removed. Thus, only proper text remains for further analysis. Case Folding is then done, which makes all words in lower case so that any variation that could be caused due to different cases, such as words in uppercase and lowercase, is eliminated. Tokenization is then performed to divide the text into smaller elements, normally words or extremely short sentences, in order that the textual material can be analyzed more precisely. This is then accompanied by Stopword Elimination, which eliminates popular but semantically void words such as "the," "and," or "is." These words are common but do not contribute significantly to sentiment analysis. Normalization is next to standardize variations in spelling, slangs, or abbreviations so that the model can better process the text. Finally, Stemming is done, which reduces words to base form. For example, words like "running" or "ran" would both be reduced to base word "run." This is done to aggregate various inflections of a word into one feature so that the model gets more efficient.

Once the data has been preprocessed appropriately, it is divided into Training and Test datasets during the Data Splitting phase. This division ensures that the model is trained on one subset of the data and tested on another subset, providing an unbiased measure of how well it will generalize to new, unseen data. The Training Data is used to train the model in sentiment classification based on the reviews and what they include, and the Test Data is reserved for model testing to study how well the model does with unseen data.

The essence of the study happens in the Model Training phase when the CNN-LSTM is employed. The CNN component of the model does a great job in extracting local spatial features from the text, identifying patterns and significant phrases out of the reviews. However, with a single CNN, it's impossible to assist in capturing long-term dependencies in the text, and that's why an LSTM layer is employed. LSTM helps the model learn and remember patterns from longer sequences of text so that it can pick up on the overall context in which words or phrases are used. By combining these two strong neural network architectures, the model can learn both local patterns and global context and hence is very effective for sentiment analysis in complex reviews.

Once training of the model is completed, the next stage is the Testing phase when the model is tried out with Test Data. This helps us conduct an impartial judgment about how effective the model would be when applying unseen data or actual use/applicability in the real world. Finally, in the Evaluation stage, performance of the model is calculated through multiple vital factors: accuracy, precision, recall, and F1-score. These provide the assistance in determining the degree to which the model predicts sentiment and aspect data, providing a well-balanced measure of correctness and completeness of the predictions.

With this systematic research process, the study seeks to acquire a deep understanding of user perception regarding Gojek services to enable action points for service improvement. Figure 1 demonstrates the flow of research, presenting phases from data collection to analysis, outlining distinctly the logical sequence of the process of sentiment analysis.

2.2. Multi-aspect Labeling

The process of labeling is a crucial step to prepare the dataset for further analysis. Labeling, in this research, is classified into two broad categories: aspect labels and sentiment labels. Both categories are needed to enable the model to understand and classify not only the general sentiment of a review, but even its specific facets—such as customer care, ease of service, or comfort—that add up to more granular insights on user satisfaction with the Gojek application.

2.2.1. Sentiment Labeling

Sentiment labeling forms the initial classification layer in this study, which is employed to categorize user reviews as negative or positive sentiments. It is an important process employed for analyzing the overall trends in sentiments of users depending on their experience with Gojek services.

In this research, sentiment labels are to be predicted based on the rating scores employed within the dataset. Ratings range from 1 to 5, with positive reviews labeled 4 or 5 and negative reviews labeled 1 or 2. Those labeled neutral (e.g., 3) are excluded to maintain clear polarity in the sentiment label. It is thus possible to have consistency and interpretability of the sentiment labels directly taken from user-provided ratings.

Sentiment breakdown in the dataset emerges in Table 1, with 14598 positive and 3973 negative reviews. This two-way sentiment classification makes it easy to conduct the following analysis of public opinion and determine key strengths and weaknesses of Gojek's services.

Table 1. Sentiment Label Distribution in the Dataset

Sentiment	Number of Reviews
Positive	14598
Negative	3973

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

2.2.2. Aspect Labeling

Aspect labeling comes after sentiment labeling, where the reviews are marked according to what precisely regarding the Gojek service is being said. More specific than sentiment labeling, aspect labeling allows for more precise details of what elements of the service are most important to users to be collected.

The dimensions explored here are grounded in prior research on service quality measurement and customer feedback analysis. According to UNE-EN 13816 [13], a global standard of service quality, are essential building blocks of the Gojek experience. These include Access, Information, Customer Service, Comfort, Time, Availability, and Safety, all distinct building blocks of the Gojek service.

To obtain accurate and coherent word-level labeling of them, this current study employs the Cosine Similarity-based approach, a robust tool for quantifying word similarity on a semantic, not literal, basis. Specifically, every word in the reviews is mapped to a numerical vector based on the FastText word embedding model. FastText is helpful as it generates word embeddings that both cover syntactic and semantic features of words, which is critical to address the variations in user language. For example, the term "driver" can be associated with Availability, while "fast" can be related to Time, even though the precise words do not appear in the keyword list of the aspect.

The cosine similarity method measures the similarity between vector representation of the words in the review and aspect keywords. If the similarity measure between the vector of any word in the review and aspect keyword vector is more than some fixed threshold value, then the review falls under that aspect. This approach makes the system flexible and effective as it does not require a direct one-to-one keyword match but instead takes into consideration semantic relationships between words. This ensures that reviews with the same content are categorized under the corresponding aspects even when the user uses synonyms or very closely related terms not in the given keyword list.

For example, a review containing the term "system" or "platform" can be tagged under the Access aspect, even if these terms are not an exact match for the keyword "access." The ability to accommodate such flexibility enhances the capability of the system to detect and tag subtle fine-grained details in user feedback, and hence make the aspect labeling robust.

This classifying by aspects method draws on previous work of Ghadery et al. (2019), where soft cosine similarity was used to find aspect categories within text without needing direct supervised labels. The method considers the semantic between words within a higher dimensional space, so even strongly related words are labeled with the appropriate aspect, hence ideal for use in this study context [14]. The recognized features and their respective keywords are presented in Table 2, and each feature is a prominent dimension of the Gojek user experience. The features allow the model to ascertain which specific service features are most impactful on user sentiment.

Table 2. Aspect Definitions and Example Keywords

Aspect	Definition	Example Keywords
Access	Reviews describing the ease of users accessing Gojek services	error, application, system
Information	Reviews containing service-related information	tariff, cost, expensive
Customer Service	Reviews referring to the ease of interaction with customer support	admin, complaint, promo
Comfort	Reviews related to user comfort during service use	helmet, air conditioning, comfortable
Time	Reviews concerning the timeliness of the service	long, fast, wait
Availability	Reviews referring to service or driver availability	driver, empty, unavailable
Safety	Reviews discussing user safety during service use	safe, accident, speeding

Upon the completion of the aspect labeling, the dataset is now ready to be trained using the labeled data. Overall, the dataset contains varying frequencies of reviews for each aspect. Access is the most named aspect with 8926 reviews, and Safety is the least named aspect with only 114 mentions as seen from Table 3.

Table 3. Aspect Label Distribution in the Dataset

Aspect	Sentiment		Number of Reviews
	Positive	Negative	
Access	6968	1958	8926
Information	1179	417	1596
Customer Service	890	289	1179
Comfort	2238	410	2648
Time	2563	567	3130
Availability	569	255	824
Safety	68	46	114
Others	123	31	154

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

Such aspect-tagged data is the precondition for the model training of CNN-LSTM, and the model, aspired to capture not only global (spatial) but also the long-range (temporal) dependencies in texts, can also predict not merely general sentiment for Gojek services but just which service characteristics influence users.

Generally speaking, the use of sentiment and aspect labeling in this research makes it possible for a more extensive analysis of users' opinions. It enables the model to generate insights that not only stay within the scope of overall sentiment but also incorporate an extensive knowledge of how users perceive various aspects of the Gojek service.

2.3. CNN-LSTM Modeling

In this research study, a combination of deep learning architecture that pairs Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) is employed in performing multi-aspect sentiment analysis on Gojek app user reviews. A hybrid framework like this is employed in light of complementary strengths: CNN excels at extracting local spatial features from the text data such as word n-grams, whereas LSTM is equally good at modeling long-distance dependencies and maintaining the sequential pattern of language intact [15] [16] [17].

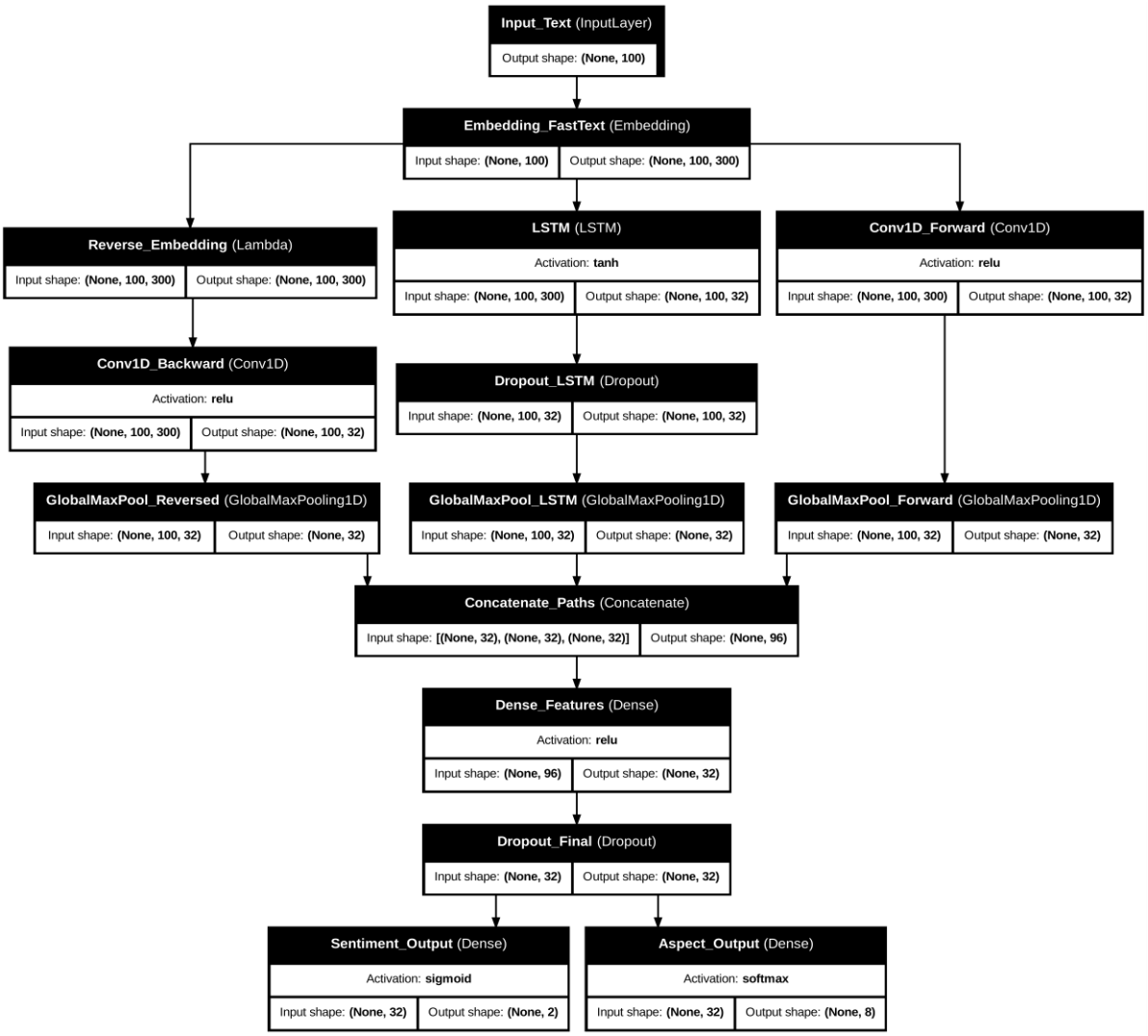

Figure 2. CNN-LSTM Architecture Used in Research

The overall design of the CNN-LSTM model used in the present study is presented in Figure 2 and illustrates processing of the input text in terms of multiple parallel branches prior to processing at the classification level. The input layer is provided with tokenized and padded sequences of the text resulting from the preprocessing operation. The tokens are then mapped to dense vector representations using an application of an embedding layer, which utilizes pre-trained FastText embeddings. Notably, the embeddings are not trainable at training time, so the model can leverage semantic knowledge acquired from a large external corpus without altering them in task-specific learning [18].

The model architecture comprises two one-dimensional convolutional paths intended to extract spatial patterns from input and reversed input sequences. This two-path convolutional architecture allows the model to comprehend linguistic features that can be presented differently depending on the reading direction—forward or backward—thus improving contextual representation [19].

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

In addition to the CNN streams, a dedicated LSTM branch is used for capturing temporal and contextually conditioned dependence along the sequence length. There is also an inclusion of a dropout layer within the branch that randomly discards a subset of neurons during training for avoiding overfitting [20].

After feature extraction, each of the three branches' outputs—CNN forward, CNN reversed, and LSTM—are input to a GlobalMaxPooling1D layer, which selects the most representative features in the time direction. The vectors extracted from the above three branches are concatenated in a concatenation layer to obtain an overall feature representation that integrates the spatial and sequential information.

This combined embedding is then fed through a dense layer activated by ReLU, then another dropout layer for regularization. This architecture is finalized using a two-head output layer: one is a binary sentiment classification head, determined either as negative or positive, with sigmoid activation; the other is a multi-class aspect classification head, with softmax activation, which categorizes every review to one of the given service aspects, i.e., access, time, availability, safety, information, customer service, and comfort.

This double-headed output design plays a central role in enabling the model to learn and predict simultaneously both sentiment and applicable service dimensions from any user review—rendering it well-suited for multi-aspect sentiment analysis tasks that require a high level of granularity to understand user sentiment across multiple elements of service [21] [18].

3. RESULT AND DISCUSSION

This study produced a Convolutional Neural Network – Long Short Term Memory (CNN-LSTM) model that was trained to perform two primary tasks: sentiment classification and aspect extraction on Gojek user reviews. The model was trained on a dataset containing 18571 customer reviews from the Google Play Store platform. This information was split into two: train set data and test set data, in accordance with the typical machine learning training-test split to enable the model to generalize to future data. On completion of the model training, the test using the test data was performed to validate the performance of the model in performing the two main tasks:

a. Sentiment Classification
   The model properly identified the sentiment of the user reviews as either positive or negative. Sentiment labels were given depending on rating scores by the users per review. More specifically, reviews with ratings 4 and 5 were given a positive label, and ratings 1 and 2 were given a negative label. Reviews with a neutral rating (i.e., 3) were excluded to maintain the polarity and clarity of sentiment contrast. This rule-based labeling approach enabled the model to learn from users' explicit evaluations directly, projecting sentiment classification onto real-world user satisfaction measures. The model exhibited outstanding performance in predicting the sentiment orientation of users' reviews, thereby providing insightful information about public attitudes towards Gojek's services.
b. Aspect Classification
   In addition to sentiment labeling, the model was also trained to detect the service areas being complained about in customer reviews. The service areas include access, time, availability, safety, information, customer service, and comfort. Aspect labeling was performed using a cosine similarity-based approach, which was augmented by FastText embedding-based word representations, to enable the model to project words in the reviews onto pre-defined aspect keywords. This method allowed the model to accurately categorize text into the relevant service categories even when the exact aspect terms were not stated.

The performance of the model was measured using accuracy as the primary metric for both sentiment and aspect classification. The performance outcomes showed how well the model performed in extracting and analyzing user opinions on various aspects of the Gojek service.

3.1. Data Exploration

Before delving into the modeling process, it is important to gain a preliminary understanding of the linguistic characteristics of the dataset. For this reason, wordclouds were built for each one of the seven service dimensions: Access, Time, Comfort, Information, Customer Services, Availability, and Safety. These visualizations provide an intuitive image of the most frequent words used in user complaints concerning each dimension. By highlighting key words and themes, word clouds give us useful information regarding the major themes and sentiment indicators in the data and hence influence subsequent preprocessing and model architecture choices.

(a) Access Aspect

(b) Availability Aspect

(c) Comfort Aspect

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

(d) Customer Services Aspect (e) Information Aspect (f) Safety Aspect

(g) Time Aspect

Figure 3. Word cloud visualization of prominent terms in user feedback related to the Aspects

The Access word cloud is one of positive sentiment as well, with encouraging words like "sangat membantu," "terima kasih," and "mantap" being employed in the expression of satisfaction with the convenience of the app. Salient mentions of "gojek" and "aplikasi ini" highlight ease of use and simplicity of the mobile app. However, terms like "enggak bisa" and "susah" suggest that access issues exist among some users, likely due to flaws in their devices or capacity limitations of the service. References to promotions ("banyak promo," "diskon") and driver availability ("dapat driver") suggest that access also encompasses responsiveness of service and incentive reach. Overall, users appreciate the ease of Gojek's access, although dependability in operation and greater availability of services are still basic expectations.

User reviews on Availability also point to the value of the Gojek app and its services like GoFood, with extremely common instances of "aplikasi," "gojek," and "gofood." Adjectives like "sangat membantu," "mantap," and "bagus" suggest that the app is reliable for transportation and food delivery. Phrases like "pakai," "pesan gofood," and "saldo gopay" suggest frictionless transactional usage. But in spite of this, technical challenges are still evidenced by such terms as "susah," "lemot," "error," and "blokir," indicative of technical limitations. Despite these, however, the majority of customers also highlight how the app "memudahkan" and "mempermudah" daily activities, indicative of excellent satisfaction with access, qualified by a call for higher stability and less technological issues.

For the Comfort domain, users highlight emotional and physical comfort while utilizing Gojek. Words like "sangat membantu," "nyaman," "aman," and "menyenangkan" show that riders are comfortable and feel safe on trips. Driver attitude is also significant, with several uses of "driver," "ramah," and "sopan" showing how courteous service enhances comfort. Pleasant phrases also include "cepat," "lancar," and "bagus," showing that speed contributes to a smooth ride. While there are some who say "sulit" or "ribet," overall the feeling is overwhelmingly positive, indicating that Gojek has succeeded in providing a pleasant, easy-to-use service.

Customer feedback on customer service is dominated by offers, as there are ubiquitous uses of "promo," "voucher," and "diskon," indicating that the user is motivated by low prices. Phrases like "terima kasih" and "mantap" indicate gratitude, but cries like "tolong," and complaints like "enggak ada," or "pelit" bear witness to frustration when promos are scarce or vague. Tests for "respon cepat" and "cs" affirm users prefer rapid support, though labels like "komplain" and "blokir" affirm inconstancies in service. Although there is gratitude for Gojek's monetary rewards and assistance, a clear demand for more consistent communication and more promotional access exists.

Information dimension shows that users associate Gojek's ease and app convenience with quality experience. Adjectives like "pelayanan," "aplikasi," and "layanan" highlight satisfaction with the delivery of information. Positives are "bagus," "ramah," and "terima kasih," especially for useful functionality and navigability. Fears about "harga," "ongkir," and "bayar" show that users want more transparent pricing. Negatives like "enggak jelas" and "error" reflect momentary confusion or bugs. Overall, users value timely and concise information but want enhanced cost transparency and frequent updates.

Safety feedback is about account security, driver behavior, and payment consistency. The prevalence of high-frequency words "driver," "akun," "verifikasi," and "blokir" demonstrates users' concerns for login issues and verification issues, which are typically described as difficult. Referrals to "gopay," "limit," and "saldo" show trust in online payments is necessary to feel safe. However, the users appreciate through phrases like "aman," "bagus," and "terima kasih" when the operations go smoothly. Emotional cues like "tolong" and "semoga" convey frustration as

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

well as hope. Users are aware that there is a need for safety features but do not want rougher and more complicated procedures.

Customers usually commend Gojek's timeliness and speed, with common descriptors like "tepat waktu," "cepat," and "sangat membantu" conveying satisfaction in day-to-day language. Phrases like "cepat sampai" and "selalu tepat" attest to appreciation for timely arrivals. Usage of "lama" and "lama banget" attests that occasional delays occur. Positive adjectives like "mantap," "baik," and "terima kasih" reinforce Gojek as a time-saving alternative. Despite trivial grouses, customers view Gojek as efficient and reliable for on-call mobility.

The word cloud findings in each category present a graphical and data-based visualization of the experience and sentiments of users towards Gojek services. The repeated positive terms like "sangat membantu," "mantap," "bagus," and "terima kasih" reveal high overall satisfaction, particularly with ease of access, consistency, and interaction with the driver. At the same time, the prevalence of concern words such as "blokir," "susah," "lama," and "error" points to those areas of priority in which service delivery can be enhanced. These patterns of images are placed to highlight the dual nature of user sentiment—overwhelmingly positive but peppered with points of pain. By taking into account the salience and the context of these keywords, the word clouds provide important insights into where Gojek is strong and where most improvement is required to make it an even more seamless, reliable, and fulfilling user experience.

3.2. Experimental Description

The experiment was designed to test the performance of the CNN-LSTM model in simultaneous sentiment and aspect classification under multi-task learning setting. The approach is better because it enables the model to capture local patterns as well as long-distance dependencies of the text data using the power of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM).

To evaluate the effect of model architecture on its performance, the three test cases were run with varying numbers of units between layers: 32, 64, and 128 units. Each group of configurations was used in a similar model for the purpose of obtaining a similar evaluation on the effect of classification accuracy introduced by different sizes of different units. This test setup was required to find out how changes in model complexities influenced the classification operations and if including more neurons resulted in improved performance or brought forth any possible overfitting. Results from these different setups are covered in following sections, which allow us to determine the best configurations for the CNN-LSTM model.

3.3. Classification Results

The performance of the CNN-LSTM model was further investigated by determining its ability to classify user sentiment—positive and negative—on seven pre-defined service areas: Access, Time, Comfort, Information, Customer Services, Availability, and Safety. Classification results in each of the areas in three differently sized hidden layers (32, 64, and 128 units) are presented in Table 4. The results provide detailed information regarding how the complexity of the model influences sentiment detection accuracy across different categories of user feedback.

Table 4. Sentiment classification accuracy across aspects and model configurations

Aspect	Sentiment					
	Negative (%)			Positive (%)		
	32 Layer	64 Layer	128 Layer	32 Layer	64 Layer	128 Layer
Access	76.7612	78.5504	74.4634	93.7491	93.5205	93.6423
Time	74.0894	74.9775	74.2664	95.5069	95.6704	95.0978
Comfort	64.878	64.6341	69.5122	95.2558	95.1163	94.5581
Information	69.0476	70.4877	72.1658	91.0213	91.2916	90.3935
Customer Services	71.954	75.4325	72.6074	93.6213	92.3251	92.7971
Availability	72.9412	71.3725	70.1961	89.1796	89.7439	90.3046
Safety	60.8889	65.1111	58.6667	75.4545	71.8182	71.9697

As seen from Table 4, the overall model performed well in detecting positive sentiment with very high accuracy for all aspect classes. Time and Comfort aspects worked best alone in positive sentiment tagging—95.67% for the 64-unit case of the Time aspect. This uniformity causes the users to be more specific and consistent in their positive comments for these service attributes so that the model can better learn about positive sentiment patterns.

Negative class classification, on the other hand, was closer to accuracy in aspect and in configuration. In Access, for example, the highest in negative accuracy was 78.55% with the 64-unit model but fell off to 74.46% when units were doubled to 128. Similarly, in Customer Service, the highest in negative accuracy was the 64-unit model at 75.43%, then followed by the 32-unit model, then the 128-unit model. These findings suggest that augmenting the number of neurons may enhance capability but perhaps add redundancy or overfitting that degrades generalization, especially for low-frequency or context-dependent negative sentiment.

The Safety category proved to be the most challenging for the model, with the lowest overall accuracy, particularly for negative sentiment. The best performing configuration (64 units) produced only 65.11% accuracy, which indicates that customer grievances against safety might be more varied and contextually informed. Such words

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

might be tied to fine-grained linguistic cues that even the model, with LSTM's temporal learning capacity, could not reliably identify.

Most notably, the Information topic exhibited most apparent augmented negative sentiment prediction with deeper model depth. Accuracy went up from 69.05% to 32 units to 72.17% to 128 units, demonstrating that this topic benefited from greater semantic modeling depth. This benefit was at a minimal sacrifice of positive sentiment accuracy from 91.29% to 64 units to 90.39% to 128 units, further supporting the sacrifice typically achieved in deep learning architectures of accuracy vs. overgeneralization.

In general, the 64-unit model was the best balanced and effective architecture. It produced satisfactory performance for sentiment polarities as well as for all aspect types overall without the effect of diminishing returns and overfitting risk seen in the 128-unit architecture. While the 32-unit model trained faster and had satisfactory results for straightforward aspects like Availability, its overall performance was mostly below optimal, particularly in detecting complex negative sentiments.

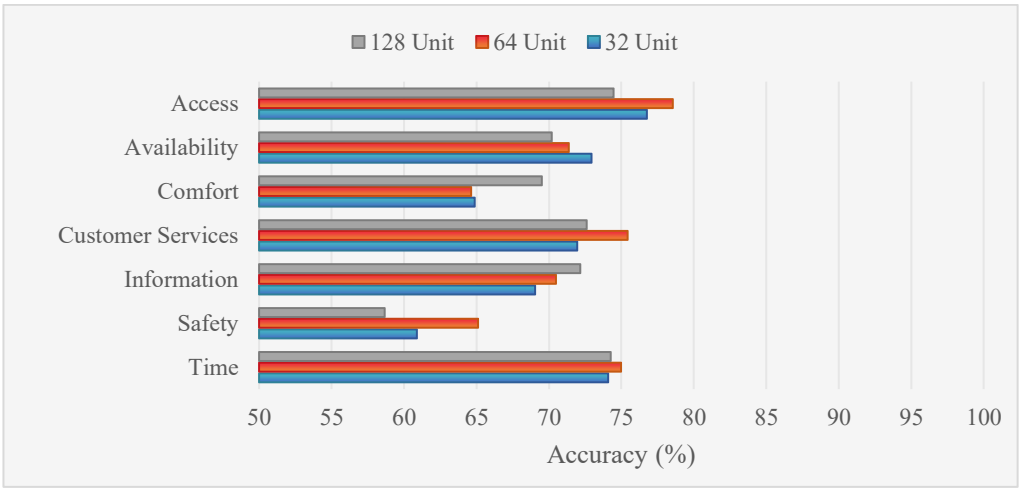

Figure 4. Negative Sentiment classification accuracy across Aspects within different model configurations.

Figure 4 displays the classification accuracy of the CNN-LSTM model for negative sentiment across different areas of services under three configurations of hidden layers—32, 64, and 128 units. The clear trend from the figure is the persistently better overall performance of the 64-unit configuration, which outperforms the 32 and 128-unit models in the majority of areas. Interestingly enough, Customer Services and Access were both highest in negative sentiment accuracy with 64 units at 75.43% and 78.55%, respectively. Conversely, the 128-unit configuration often reported reducing accuracy, suggesting potential overfitting or reduced generalization capability when complexity in the model is greater. The Safety component, the most challenging category, maintained accuracy at an abysmal 65.11% at best with 64 units, capturing the most challenging it is to get consistent sentiment classification. Notably, the Information facet followed a unique increasing trend, systematically increasing from 69.05% (32 units) to 72.17% (128 units), indicating that lower-level models are perhaps better suited for semantically rich and complex classes. This visual analysis reaffirms the earlier interpretation that the increasing of models does not necessarily lead to improvement, and that well-configured combinations (such as 64 units) perform best at capturing the diverse expressions of negative sentiment across different service domains.

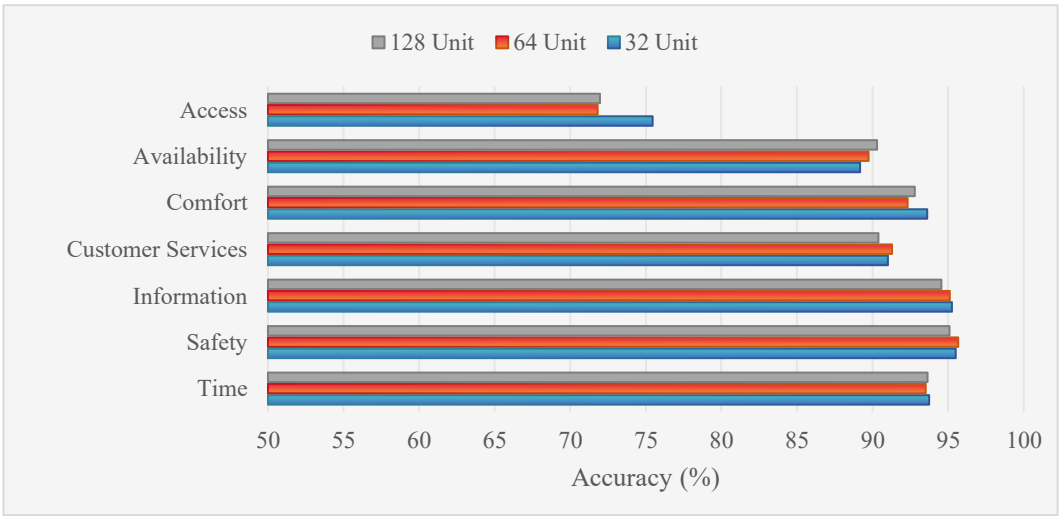

Figure 5. Positive Sentiment classification accuracy across Aspects within different model configurations.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

Figure 5 displays the accuracy of the CNN-LSTM model in classifying positive sentiment over all the aspects of service under varying configurations of hidden units. The overall performance is high across the board, with the accuracies being more than 90% for all the categories almost everywhere. The Time aspect shows outstanding stability over configurations with an accuracy of more than 93.5% everywhere. Similarly, Safety and Information dimensions show minimal fluctuation, with best scores of 95.67% and 95.26%, respectively, suggesting that such positive feedback in these dimensions is stable in terms of language and easily attainable for the model. Unlike negative sentiment, the 128-unit model did not lead to performance degradation in the majority of cases, and improved accuracy ever so slightly in Availability, from 89.18% to 90.30%. But Access varies with considerably less accuracy than other characteristics—dropping from 75.45% (32 units) to 71.82% (64 units) and 71.97% (128 units). That suggests that in positive feedback, Access is more vague or subtle, which is harder for the model to learn from. In general, the graph corroborates that positive sentiment is easier to classify, or maybe because users are more repetitive and patterned in their language when satisfied.

These classification results emphasize the key observation: user dissatisfaction utterances are usually more linguistically diverse and harder to classify, requiring more generalization of higher-level, more variable patterns by the model. Positive sentiment, however, is usually more formulaic and predictable, which might be the explanation for the consistently good performance on most metrics and setups. Explanation also backs up that model capacity should be well-tuned since increasing it does not always improve performance and can even harm model generalizability to some tasks.

3.4. Discussion

The classification results provide useful feedback on the performance dynamics of the CNN-LSTM model in the context of being trained on user-generated Gojek reviews with a text-to-sequence input pipeline combined with 300-dimensional FastText word embeddings. The embedding approach allowed the model to discover semantic relationships and contextual similarity between words, enriching the input representation with pretrained linguistic knowledge. As a result, even lexically different but semantically similar words (for example, "lagging" and "slow") may be projected to nearby vector spaces, which facilitates the model's ability to recognize sentiment and aspect relevance in varied language.

Among the tested model, the 64-unit configuration consistently yielded the best performance, striking an optimal balance between model complexity and generalization ability. As shown in Table 4, the 64-unit model achieved the highest or near-highest accuracy for both positive and negative sentiment classification across most service aspects, including Access, Time, Customer Services, and Information. This configuration offered sufficient representational capacity to extract and model important patterns in the embedded text sequences while avoiding the pitfalls of overfitting.

The FastText embedding was one of the main drivers of the model's performance. With every word being represented as a 300-dimensional vector, the model received dense, subword-aware representations that preserved syntactic and semantic relationships even in out-of-vocabulary words or typo words common in app reviews. The embeddings were particularly effective in aspect classification as they enabled the model to infer aspect relevance even when there were no keyword matches in the reviews explicitly. The CNN layers complemented this by pulling out local n-gram patterns from such embeddings, whereas the LSTM layers modeled the sequential dependence and long-range contextual cues, both of which are crucial for representing subtle sentiment expressions.

The superior performance of the 64-unit model is particularly notable when compared to the other two configurations. The 32-unit configuration, though computationally efficient, frequently underperformed in negative sentiment detection—likely due to underfitting. Its limited capacity was insufficient to model the diversity and subtlety of user complaints, especially in aspects like Comfort and Safety, which often involved less explicit or emotionally complex language.

In contrast, the 128-unit setup, while sometimes equaling or surpassing the 64-unit performance on some negative sentiment tasks (e.g., Information, Comfort), showed signs of overfitting. This configuration likely picked up too much noise in the training data, which hurt its generalization to new examples. In particular, it suffered from a drop in positive sentiment classification in some categories (e.g., Information, which dropped to 90.39%) with increased model capacity, which is evidence of the risk of diminishing returns with larger architectures.

Further, the model's consistently higher accuracy in positive sentiment classification across all categories and configurations captures the simpler, more regular language generally used to convey satisfaction. Positive sentiment will tend to feature standard patterns and stronger sentiment cues, which are more easily detected by both CNN filters and LSTM memory cells. Negative sentiment is more linguistically diverse, being more likely to contain sarcasm, indirect insults, or contextual references that expend effort on both lexical and sequential models.

These results validate the necessity of quality embedding and model tuning. The use of 300-dimensional FastText embeddings, which were trained on massive corpora and could learn subword-level meaning, provided rich input representation that allowed the CNN-LSTM architecture to perform well even with moderate model complexity. The 64-unit model exploited the rich input without overfitting, thereby being the best configuration for this classification task.

In summary, the integration of text-to-sequence encoding and 300-dimensional FastText embeddings with a 64-unit CNN-LSTM architecture permitted accurate classification of sentiment and service dimensions in noisy and

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

unstructured user review data. The study underlines the necessity for balancing model complexity with the richness of input representation and with the inherent variability of the classification problems. The findings also suggest that further gains can be achieved through the application of more advanced contextual embedding techniques, e.g., transformer-based models, or with the addition of attention mechanisms, which potentially capture the finer and contextually sensitive nature of sentiment—particularly in such subtle domains as safety. Such gains could lead to greater model robustness and transparency when deployed in real-world sentiment analysis applications.

4. CONCLUSION

This study developed and validated a CNN-LSTM model that aims to classify sentiment and extract aspects at the same time from Gojek user feedback. Based on a text-to-sequence input and 300-dimensional FastText embeddings, the said model was capable of deriving semantic relationships and context dependencies from noisy user-written text data. Experimental results indicated that the 64-unit configuration of the CNN-LSTM model provided the best trade-off between model complexity and generalizability, with the best accuracy across most sentiment and aspect classes. The improved performance of the model in positive sentiment classification and its ability to handle complicated and diverse negative feedback confirm the strength of combining pretrained embeddings with a hybrid CNN-LSTM structure. The results suggest that precise adjustment of model capacity is required to avoid underfitting or overfitting when faced with complex natural language tasks. In short, the proposed CNN-LSTM model is a robust approach to summarizing useful insights from large volumes of customer opinions to provide useful information that can be used to improve service quality and user satisfaction in online platforms like Gojek. Future work can explore the addition of more advanced embedding techniques, i.e., transformer models or attention mechanisms, to better learn nuanced and context-dependent sentiment nuances via features such as safety. These additions can possibly improve the accuracy and interpretability of sentiment systems in real-world applications.

REFERENCES

[1] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, “Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm,” *2021 Int. Semin. Mach. Learn. Optim. Data Sci. ISMODE 2021*, pp. 163–167, 2022, doi: 10.1109/ISMODE53584.2022.9742906.

[2] C. A. Haryani, A. E. Widjaja, H. Hery, and F. V. Ferdinand, “Sentiment Analysis of User Satisfaction Towards Sales Promotion of Gojek Application Service Using Support Vector Machine (SVM),” *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 14, no. 2, pp. 66–70, 2023, doi: 10.31937/si.v14i2.3398.

[3] A. Ramadina, K. Tania, A. Wedhasmara, and et al., “Knowledge Extraction of Gojek Application Review Using Aspect-based Sentiment Analysis,” *Indones. J. Comput. Sci.*, vol. 13, pp. 3962–3976, 2024, doi: 10.33022/ijcs.v13i3.4020.

[4] P. Kurniawati, R. Y. Fa’rifah, and D. Witarsyah, “Sentiment Analysis of Maxim Online Transportation App Reviews using Support Vector Machine (SVM) Algorithm,” *Build. Informatics, Technol. Sci.*, vol. 5, no. 2, pp. 466–475, 2023, doi: 10.47065/bits.v5i2.4265.

[5] A. Nur, A. Zulkifli, and N. A. Shafie, “Review of the Lazada application on Google Play Store: Sentiment Analysis,” *J. Comput. Res. Innov.*, vol. 9, no. 1, p. 2024, 2024, doi: 10.24191/jcrinn.v9i1.412.

[6] A. T. Rizkya, R. Rianto, and A. I. Gufroni, “Implementation of the Naive Bayes Classifier for Sentiment Analysis of Shopee E-Commerce Application Review Data on the Google Play Store,” *Int. J. Appl. Inf. Syst. Informatics*, vol. 1, no. 1, pp. 31–37, 2023, doi: 10.37058/jaisi.v1i1.8993.

[7] M. Fahmi, A. Puspita, and Y. Yuningsih, “Sentiment Analysis of Online Gojek Transportation Services on Twitter Using the Naïve Bayes Method,” *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 8, no. 2, pp. 84–90, 2023, doi: 10.33480/jitk.v8i2.4004.

[8] A. Fathrizqy, P. Mahardika, and A. Muid, “The Sentiment Analysis of Online Customer Review on Food and Beverages Delivery Services in the GOJEK Application Using K-Nearest Neighbors,” Proceedings of the 3rd Asia Pacific International Conference on Industrial Engineering and Operations Management, pp. 854–862, 2023, doi: 10.46254/ap03.20220167.

[9] R. A. A. Malik and Y. Sibaroni, “Multi-aspect Sentiment Analysis of Tiktok Application Usage Using FasText Feature Expansion and CNN Method,” *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 277–285, 2022, doi: 10.47065/josyc.v3i4.2033.

[10] D. Wahyudi and Y. Sibaroni, “Deep Learning for Multi-Aspect Sentiment Analysis of TikTok App using the RNN-LSTM Method,” *Build. Informatics, Technol. Sci.*, vol. 4, no. 1, pp. 169–177, 2022, doi: 10.47065/bits.v4i1.1665.

[11] Syaiful Imron, E. I. Setiawan, Joan Santoso, and Mauridhi Hery Purnomo, “Aspect Based Sentiment Analysis Marketplace Product Reviews Using BERT, LSTM, and CNN,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 3, pp. 586–591, 2023, doi: 10.29207/resti.v7i3.4751.

[12] A. M. Alayba and V. Palade, “Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9710–9722, 2022, doi: 10.1016/j.jksuci.2021.12.004.

[13] Annisa Octaviana Sari, “Multi-Aspect Sentiment Analysis pada Jasa Layanan Transportasi Online Menggunakan Metode Maximum Entropy (Studi Kasus: Go-Jek dan Grab),” 2019.

[14] E. Ghadery, S. Movahedi, H. Faili, and A. Shakery, “An Unsupervised Approach for Aspect Category Detection Using Soft Cosine Similarity Measure,” arXiv, December 2018, 2018, doi: 10.48550/arXiv.1812.03361.

[15] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Springer International Publishing,* vol. 8, no. 1., 2021. doi: 10.1186/s40537-021-00444-8.

[16] F. Aksan, Y. Li, and V. Suresh, “CNN-LSTM vs . LSTM-CNN to Predict Power Flow Direction :,” 2023.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 696−707
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7511

[17] X. Wen and W. Li, "Time Series Prediction Based on LSTM-Attention-LSTM Model," *IEEE Access*, vol. 11, no. April, pp. 48322–48331, 2023, doi: 10.1109/ACCESS.2023.3276628.
[18] V. No, N. El Koufi, Y. M. Missah, and A. Belangour, "A Hybrid CNN-LSTM Based Natural Language Processing Model for Sentiment Analysis of Customer Product Reviews : A Case Study from Ghana," *Journal of Hunan University Natural Sciences*, vol. 51, no. 8, 2024, doi: 10.55463/issn.1674-2974.51.8.5.
[19] Firmansyah, D. P. Rini, and Sukemi, "Klasifikasi Data Penderita Skizofrenia Menggunakan CNN-LSTM dan Cnn-Gru pada Data Sinyal EEG 2D," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 7, no. 4, pp. 642–650, 2023, doi: 10.35870/jtik.v7i4.1072.
[20] D. O. Oyewola, L. A. Oladimeji, S. O. Julius, L. B. Kachalla, and E. G. Dada, "Optimizing sentiment analysis of Nigerian 2023 presidential election using two-stage residual long short term memory," *Heliyon*, vol. 9, no. 4, p. e14836, 2023, doi: 10.1016/j.heliyon.2023.e14836.
[21] C. N. Dang, M. N. Moreno-García, and F. De La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9986920.