Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Evaluating User Satisfaction in the Halodoc Application Using a Hybrid CNN-BiLTSM Model for Sentiment Analysis

Dian Kurniasari[1*], Arif Su'admaji[2], Favorisen R. Lumbanraja[3], Warsono[4]

[1,2,4]Department of Mathematics, Faculty of Mathematics and Natural Science, Lampung University
[3]Department of Computer Science, Faculty of Mathematics and Natural Science, Lampung University
[1.2,3,4] Jl. Prof. Sumantri Brojonegoro No.1, Gedong Meneng, Bandar Lampung, Indonesia

Article:
Accepted: July 10, 2025
Revised: March 02, 2025
Issued: October 30, 2025

***Correspondence Address:**
dian.kurniasari@fmipa.unila.ac.id

ABSTRACT

The growing demand for digital healthcare services in Indonesia has driven the adoption of Online Healthcare Applications (OHApps) such as Halodoc. Despite over 65 million users, maintaining user satisfaction remains a challenge. This study employs sentiment analysis using a hybrid Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) model to classify user review ratings. A dataset of 10,000 Google Play Store reviews was divided into COVID-19 and post-pandemic segments. The methodology includes data collection, pre-processing, and dataset segmentation for training, validation, and testing. Results indicate that the CNN-BiLSTM model surpasses traditional machine learning by combining CNN's feature extraction with BiLSTM's long-term dependency capture, achieving 98.71% accuracy on COVID-19 data and 98.16% post-pandemic. Additionally, the model demonstrates strong performance across other key evaluation metrics, with precision, recall, and F1-score. Misclassification analysis highlights minor errors, particularly in ratings 4 and 5. These findings help healthcare providers enhance digital services by identifying user concerns, improving platform features, and optimizing customer engagement. Beyond healthcare, this approach has real-world applications in e-commerce and financial services, where sentiment analysis informs user experience improvements.

Keywords : *hybrid CNN-BiLSTM; deep learning; text mining; halodoc.*

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

1. INTRODUCTION

The demand for both digital and non-digital health services is crucial for the broader community, particularly in Indonesia. In response to this demand, both private and public sector entities have initiated the development of computer and internet technologies, thereby transitioning healthcare services from offline to online formats, commonly referred to as Online Healthcare Applications (OHApps). Among the largest and most recognized platforms is Halodoc, which facilitates access to healthcare and pharmaceutical services via the Internet. Halodoc has been characterized in various literature as a doctor consultation application [1], [2], a health app [3], and mobile health app [4], [5], [6]. The application is accessible on both Android and iOS platforms and is available in two languages: English and Indonesian. Furthermore, Halodoc collaborates with pharmacies, hospitals, and healthcare professionals. Despite boasting over 65 million users and a valuation exceeding US$ 65 million, ensuring user satisfaction with Halodoc's services remains a significant challenge that warrants attention [7], [8].

Chatterjee [9] posits that reviews, customer ratings, and consumer recommendations serve as quantitative metrics that effectively gauge customer satisfaction via sentiment analysis. This analytical approach automates the evaluation of emotions, opinions, and sentiments expressed in written text, enabling the efficient extraction of valuable insights from extensive datasets. Such capabilities are instrumental in enhancing the understanding of customer satisfaction levels [10]. As a prominent application within the realm of Natural Language Processing (NLP) and text classification, sentiment analysis has garnered significant interest from researchers and practitioners across various domains [11], [12], [13], [14].

Text classification through Machine Learning (ML) is a widely utilized technique across numerous applications; however, this methodology presents certain limitations. A significant drawback is that ML algorithms frequently struggle to effectively extract specific features from textual data, particularly those that are distinctive and pertinent for differentiating between various classes within a dataset. This limitation arises from ML's inherent challenges in capturing the intricate patterns present in text. In contrast, Deep Learning (DL) algorithms offer a robust alternative, as they possess the capability to identify crucial features within text data autonomously. Neural networks, a prominent type of DL algorithm, excel at learning more complex patterns and interrelationships, which not only facilitates the identification of distinguishing characteristics but also enhances the accuracy of text classification efforts [15].

Several DL methodologies have demonstrated effectiveness in the classification of textual data, including Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM) networks. CNNs are multilayered neural architectures specifically engineered based on the operational principles of biological visual systems [16]. In the context of textual analysis, CNNs are employed to extract salient features that encapsulate critical attributes of the text [17]. This technique facilitates the identification of pertinent patterns and structures within the text, thereby streamlining the classification process.

LSTM networks are an advanced variant of Recurrent Neural Networks (RNNs) specifically engineered to mitigate the limitations of RNNs in effectively storing and utilizing long-term information [18]. LSTMs employ a mechanism known as "gates" which determines which pieces of information should be retained in memory and which should be discarded [19]. This capability enhances the model's effectiveness in analyzing long-term dependencies within text data, thereby enabling it to preserve essential contextual information from preceding data crucial for the classification task. Nonetheless, a significant drawback of conventional LSTMs is their restriction to processing only historical sequences, rendering them incapable of incorporating potentially relevant future information. This limitation is addressed through the implementation of Bidirectional Long Short-Term Memory (BiLSTM) networks. BiLSTMs are designed to process information from both past and future contexts, allowing for a more comprehensive understanding of the overall context within a given sentence or data sequence [20].

Despite the effectiveness of various previously discussed DL methodologies in text

data classification, several research gaps remain to be explored. Primarily, while CNNs demonstrate superior performance in feature extraction and pattern recognition, they cannot capture long-term dependencies within data sequences. Conversely, LSTM networks and BiLSTM networks can address this shortcoming by retaining long-term contextual information. Nonetheless, relying solely on either LSTM or BiLSTM models does not fully leverage the feature extraction strengths inherent to CNNs.

Furthermore, a significant number of existing studies primarily concentrate on employing a single model type, neglecting the potential benefits of model integration that could enhance classification performance. Consequently, there remains an opportunity to develop and investigate hybrid methodologies that amalgamate the strengths of various models, such as the combination of CNN and BiLSTM, to augment the efficacy of text classification.

This research introduces a hybrid framework that integrates CNN and BiLSTM to categorize user review ratings for the Halodoc healthcare application. This strategy aims to enhance classification accuracy by capitalizing on the feature extraction capabilities of CNNs and the long-term contextual comprehension offered by BiLSTM.

Although numerous prior studies have predominantly concentrated on employing a singular model type, a notable trend has emerged in recent years toward the development of hybrid methodologies. This innovative approach encompasses not only the amalgamation of CNN and BiLSTM models but also the integration of alternative combinations, such as CNN with LSTM, CNN with RNN, and other model pairings. The objective of this integration is to capitalize on the strengths of each model in managing data complexity, thereby enhancing accuracy and effectiveness across a range of applications.

In the domain of sentiment analysis, Ali et al. [21] introduced a classification methodology leveraging deep learning networks. Their research examined various network architectures, utilizing the Multilayer Perceptron (MLP) as a foundational framework for comparison with other models. The study incorporated LSTM and CNN, including a hybrid CNN-LSTM model, applied to an IMDB dataset comprising 50,000 movie reviews evenly split between positive and negative sentiments. Prior to analysis, the dataset underwent pre-processing using Word2Vec and word embedding techniques. The findings revealed that the CNN-LSTM hybrid model outperformed the others, achieving an accuracy of 89.2%, in contrast to CNN, which attained 87.7%, while MLP and LSTM recorded accuracies of 86.74% and 86.64%, respectively. Furthermore, the proposed deep learning model demonstrated superior performance compared to traditional approaches such as Support Vector Machine (SVM), Naïve Bayes, and Recursive Neural Tensor Networks (RNTN) discussed in prior research employing English-language datasets.

Jain et al. [22] utilized a hybrid CNN-LSTM model for the analysis of consumer sentiment. This model incorporates several advanced techniques, including dropout, max pooling, and batch normalization, to enhance its performance. The study utilized datasets from the Airline Quality database and Twitter, specifically focusing on sentiments expressed in airline reviews. They employed a hard word embedding technique that transforms textual data into numerical vectors, ensuring that words with similar meanings are positioned closely within the vector space. Furthermore, the researchers assessed the model's effectiveness by calculating various metrics, such as accuracy, precision, recall, and F1 score. The findings indicate that the proposed model outperforms traditional machine learning approaches in sentiment analysis, achieving an accuracy rate of 91.3%.

Numerous investigations into hybrid models that integrate CNN and BiLSTM architectures have been conducted, yielding promising outcomes. A notable study by Rhanoi et al. [23] introduced a hybrid model utilizing CNN and BiLSTM in conjunction with Doc2vec embeddings. This model is deemed particularly effective for sentiment analysis in lengthy texts. The research evaluated the performance of the CNN-BiLSTM model against several alternatives, including standalone CNN, LSTM, BiLSTM, and a CNN-LSTM combination, all employing Word2vec/Doc2vec embeddings. The findings revealed that the application of the Doc2vec embedding within the CNN-BiLSTM framework, specifically in the analysis of

French newspaper articles, achieved a superior accuracy of 90.66%, surpassing the performance of the other models.

Xiaoyan et al. [24] undertook a research project that introduced a sentiment analysis framework utilizing the GloVe-CNN-BiLSTM methodology, specifically designed to classify sentiments in a diverse range of text formats, encompassing both lengthy and brief entries sourced from social media platforms. In this model, GloVe is employed to create word embeddings, while CNNs are utilized to capture localized features within the text. The BiLSTM network is implemented to model temporal dependencies within the data effectively. The study leverages Twitter comments concerning COVID-19 as its experimental dataset. The findings indicate that this approach successfully identifies sentiment patterns in users' online remarks, achieving sentiment classification accuracies of 0.9565 for full texts, 0.9509 for long texts, and 0.9560 for short texts—significantly surpassing the performance of alternative deep learning models.

Ashraf et al. [25] explored the development and comparative accuracy of various integration models, specifically CNN+LSTM and CNN+BiLSTM, for the classification of music genres. Their experimental study utilized a dataset comprising 1,000 music clips, categorized into ten distinct genres—blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock—each containing 100 tracks. Each music clip has a duration of 30 seconds, a sample rate of 22,050 Hz, a 16-bit resolution for mono channels, and is encoded in MP3 format. During the pre-processing phase, the 30-second clips were segmented into 3-second intervals to enhance the accuracy of model evaluation. The results indicated that the CNN-BiLSTM hybrid model achieved an accuracy rate of 88.00%, whereas the CNN-LSTM hybrid model recorded an accuracy of 87.20%.

Singh et al. [26] introduced a novel approach to emotion analysis in Urdu, one of the primary languages in Asia. They proposed a sentiment evaluation framework for Urdu review texts by integrating BiLSTM and CNN models. The researchers conducted a comparative analysis of the performance of the CNN, LSTM, BiLSTM, and CNN-LSTM models against the CNN-BiLSTM model. The findings indicated that the CNN-BiLSTM model achieved an accuracy rate of 99%, outperforming the other models in the initial assessment. These findings align with recent studies by Ramaziyah & Setiawan [27] and Susandri et al. [28], which also confirmed the CNN-BiLSTM model's superiority over other deep learning architectures.

This research selects the CNN-BiLSTM hybrid model for its ability to combine CNN's strong feature extraction with BiLSTM's capability to capture long-term dependencies. Unlike CNN-LSTM, BiLSTM processes text bidirectionally, ensuring a deeper contextual understanding of user sentiments. This is particularly crucial for health-related reviews, where emotions and concerns are often implicit. Prior studies have shown that CNN-BiLSTM outperforms other hybrid models in text classification. Its superior accuracy makes it a practical choice for analyzing Halodoc user reviews. The insights gained can help improve digital healthcare services by addressing user concerns more effectively.

2. METHODS

This research employs a systematic approach to sentiment analysis through the implementation of a CNN-BiLSTM model. The methodology commences with the acquisition of input data, followed by an exploratory data analysis phase. These succeeded by a series of thorough pre-processing steps, which encompass case normalization, data cleansing, conjunction elimination, categorical encoding, resampling, and tokenization. The dataset is partitioned into training (85%), validation (15%), and testing (10%) subsets. Hyperparameter optimization is conducted to enhance the model's performance. The evaluation metrics applied include accuracy, recall, precision, and F1-score, thereby ensuring a comprehensive assessment of consumer sentiment.

2.1. Dataset

The dataset utilized in this study comprises review data from the Halodoc application, collected through user comments reflecting their experiences. This data was sourced from the https://play.google.com/store/apps. The data collection process involved employing the

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

Python programming language alongside the Google-play-scraper library. This approach incorporated language parameters for application reviews, specifically Indonesian (id), and targeted reviews based on their highest relevance ranking, with a total of 10,000 review entries extracted. The dataset was subsequently categorized into two distinct segments: one pertaining to the COVID-19 period and the other representing the post-COVID-19 period. Sample review data for each period is presented in Tables 1 and 2. The data consists of textual user reviews of the Halodoc health application, encompassing two variables: user reviews and ratings, which range from 1 to 5. The collected data is stored in CSV file format.

Table 1. *User review data for the halodoc application during the COVID-19 period*

Date	Rating	Review
2020-09-12	5	*Aplikasinya bagus, dokternya juga ramah pesan saya …* Eng. ver: The application functions well, and the doctor responded to my inquiries in a friendly manner …
2020-09-12	5	*Terimakasih, aplikasi yg sangat membantu dan …* Eng. ver: Thank you, this application has been incredibly helpful and …
…	…	…
2021-05-16	3	*Sedikit kecewa sih, konsul di batasi 30 mnit, baru di …* Eng. ver: I was slightly disappointed that the consultation was limited to 30 minutes, just when we were beginning to...
2021-05-16	4	*Untuk konsultasi sangat membantu, hanya saja …* Eng. ver: For consultation, it has been highly helpful, however...

Table 2. *User review data for the halodoc application after the COVID-19 period*

Date	Rating	Review
2022-05-16	5	*Sangat membantu.. Pendaftaran ke rumah sakit …* Eng. ver: It is highly beneficial... The registration process at the hospital...
2022-05-16	5	*Bener bener ngebantu sih aplikasinya juga lebih …* Eng. ver: The application has been significantly beneficial and has demonstrated enhanced capabilities…
…	…	…

Table 2 continued…

Date	Rating	Review
2023-10-17	3	*Saya puas atas info dari dokter. Terima kasih, …* Eng. ver: I am satisfied with the information provided by the doctor. Thank you.
2023-10-17	4	*Respon yang sangat cepat, …* Eng. ver: An extraordinarily quick response...

2.2. Sentiment Analysis

Sentiment analysis, often referred to as opinion mining, is a methodology designed to extract individual opinions, emotions, attitudes, and sentiments pertaining to a specific topic or circumstance from extensive collections of unstructured data [24]. With the rapid advancement of the Internet, individuals are increasingly vocal about their perspectives across diverse social platforms. These expressions encompass a broad spectrum of subjects, including online shopping experiences, reviews of recent films, political

Developments, and pressing societal issues. By systematically collecting and analyzing these comments, researchers can derive valuable insights into prevailing sentiment trends among users. This analysis not only furnishes critical information about public opinion but also serves as a powerful tool to inform decision-making across various domains, such as marketing, product innovation, and strategic communication.

Sentiment analysis represents a crucial component of Natural Language Processing (NLP), a specialized field within Artificial Intelligence that investigates the interaction between computers and human natural language [29]. Within the realm of NLP, machines acquire the ability to comprehend the syntax and semantics of human language, process that information, and generate outputs that are interpretable by users. This field encompasses the creation of computational systems capable of executing a diverse array of tasks utilizing natural language.

A vital phase in NLP is text pre-processing, which involves the cleansing and preparation of text data prior to analysis. The primary objective of this process is to transform unstructured text into a structured format that facilitates more efficient processing and

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

analysis [30]. Typically, the stages of text pre-processing comprise several critical steps, as outlined below:

a. Case folding: This procedure transforms all uppercase letters into lowercase letters within a document or converts lowercase letters to uppercase letters, thereby ensuring consistency in text processing.
b. Cleaning: At this stage, extraneous elements such as punctuation marks, emojis, and excessive whitespace are eliminated from the text, retaining only the vital information.
c. Stopword removal: Common words that do not substantially enhance the meaning of the text are eliminated. In Indonesian, examples of such stopwords include "dan" (and), "saya" (me), and "kamu" (you). This procedure can be effectively executed using the Sastrawi library in Python.
d. Tokenization: This technique divides the sentence into smaller components known as tokens, which may consist of words or sentences. Tokenization enables subsequent examination of the text's structure and semantics.

Conversely, sentiment analysis can also be seen as a form of text classification. Classification is a supervised learning technique aimed at assigning labels to documents based on predefined categories. Classification can be described as the process of mapping functions from input variables (X) to output variables (Y). Classifications are categorized by the number of possible labels, specifically binary classification and multiclass classification. Binary classification consists of two label classes, "yes" and "no," while multiclass classification involves more than two label classes [31].

2.3. Resampling

Data resampling is a strategy for addressing data imbalance in text classification. This imbalance happens when various data classes have unequal quantities, which could have an impact on the classification model's performance. The two primary resampling techniques used are oversampling and undersampling [32].

Oversampling is the process of raising the volume of data in the minority class to match that of the majority class. This strategy enhances the depiction of minority classes in the dataset. Undersampling, on the other hand, is the process of reducing the number of examples in the majority class to match the number of cases in the minority classes. This strategy decreases the majority class's dominance while increasing the model's capacity to distinguish minority classes [32]. Figure 1 shows examples of random oversampling and undersampling.

Figure 1. *Random oversampling and random undersampling*

2.4. Word Embedding

Word embedding is a NLP approach that represents words or sentences as vectors of real numbers. Word2Vec is one of the most popular ways for word embedding. This algorithm translates words to vectors based on parameters such as window size and vector dimensions. Words with comparable meanings or contexts have similar vector values and are thus clustered together in the same area of vector space. That enables Word2Vec to successfully capture word associations and similarities through large-scale corpus training.

Word2Vec offers two primary architectures for generating vector representations of words: Continuous Bag of Words (CBOW) and Skip-Gram. The Continuous Bag of Words (CBOW) architecture functions by forecasting the target

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

word utilizing the surrounding context. In contrast, the Skip-Gram model operates inversely, predicting the context based on the target word. The two architectures provide efficient solutions for various text-processing tasks, rendering them particularly appropriate for natural language analysis [29]. Figure 2 illustrates the architecture of the Continuous Bag of Words (CBOW) model and the Skip Gram model.

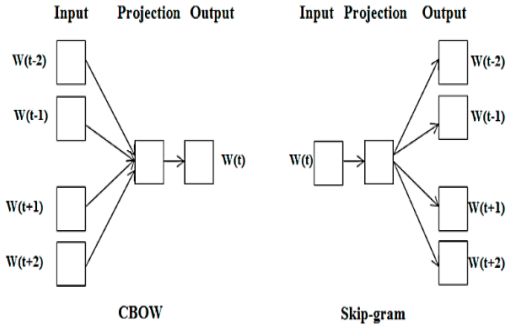

Figure 2. *CBOW and Skip Gram architecture*

2.5. Activation Function

The activation function is responsible for converting the input value into the appropriate output value. The results of this activation will be utilized as input for the next layer [33]. Activation functions that are commonly utilized include Sigmoid, Hyperbolic Tangent (TanH), Rectified Linear Unit (ReLU), and Softmax.

a. **Sigmoid**: The sigmoid function is characterized as a non-linear function that possesses differentiability and is subject to constraints regarding real input values. The derivative of this function is consistently positive. The sigmoid function is mathematically defined in Equation (1).

$$f(x) = \frac{1}{1+e^{-x}} \quad (1)$$

b. **TanH:** The main advantage of the TanH function lies in its ability to produce outputs that are centred around zero, with a value range from -1 to 1. The formula for the TanH activation function is presented in Equation (2).

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (2)$$

c. **ReLU:** The ReLU function is widely utilized as an activation function due to its superior computational speed compared to the sigmoid and tanh activation functions. The ReLU function assigns a value of zero to any input that is less than zero while it preserves input values that exceed zero. Equation (3) presents the mathematical representation of the ReLU activation function.

$$f(x) = max(0, x) \quad (3)$$

d. **Softmax:** The softmax function operates as an activation mechanism designed explicitly for multiclass classification tasks. This function produces an output that falls within the range of 0 to 1, guaranteeing that the cumulative probability sums to 1. Equation (4) defines the formula for the softmax activation function.

$$S_k = \frac{e^{y_k}}{\sum_{k=1}^{m} e^{y_k}} \quad (4)$$

2.6. Hyperparameter Tuning

The process of developing machine learning models requires the tuning of hyperparameters to determine the best combination of parameters for building an effective model [34]. The two primary hyperparameters that need to be adjusted are the learning rate and the batch size. The learning rate is a parameter that regulates the size of the modifications applied to the model's parameters with each update throughout the training process [35]. An excessively high learning rate can cause fluctuations in the model and hinder convergence, while a meagre learning rate may slow down the training process, leading to a prolonged time to achieve the desired results.

Batch size refers to the number of data samples processed simultaneously during a single iteration of parameter updates in the model training phase. The use of large batch sizes can improve the training process by optimizing parallel computing efficiency and enhancing memory utilization. A reduced batch size enables more regular and accurate updates; nonetheless, it leads to an increased training time.

Modifying hyperparameters such as learning rate and batch size is essential for enhancing model performance and facilitating efficient data learning.

2.7. Convolutional Neural Network

CNNs represent a specialized category of neural networks employed in deep learning, specifically engineered for the examination of data characterized by a structured format, including images and text [36]. The CNN model is structured with three primary components: the input layer, the hidden layer, and the output layer. The input layer is responsible for receiving the input data designated for analysis. The hidden layer consists of several sub-layers, such as the convolution layer, the pooling layer, and the fully connected layer, which operate sequentially [37].

The convolution layer performs feature extraction from the input data by applying a filter or kernel that moves across the data. The pooling layer serves to reduce the dimensions of the data, effectively lowering the complexity of calculations while maintaining critical information. The fully connected layer connects every neuron from the previous layer to the neurons in the output layer, resulting in the model's ultimate decision. This organized framework allows CNNs to acquire and recognize intricate patterns within data, establishing them as a prominent choice for a range of pattern recognition and image analysis tasks. Figure 3 presents the architecture of the CNN, as outlined in the following sections.

2.8. Bidirectional Long Short-Term Memory

BiLSTM is an enhancement of the LSTM network specifically designed to improve performance in sequence classification tasks [38]. The BiLSTM architecture, unlike the conventional LSTM model, is designed to process information by utilizing contexts from both the past and the future rather than being restricted to preceding contexts alone. This functionality is accomplished by concurrently training two LSTM networks: one is configured to process the input sequence in a forward direction from start to finish, while the other is set up to process the sequence in a backward direction from end to start. As a result, BiLSTM captures a more comprehensive context, thereby improving the accuracy of predicting the next step in the sequence. Figure 4 presents the architecture of the BiLSTM model, emphasizing the interaction between the two hidden layers.

2.9. Hybrid CNN-BiLSTM Model

The integration of the CNN and BiLSTM models aims to establish a hybrid model that leverages the strengths of both architectural frameworks. The hybrid model is structured to efficiently capture and extract features through the use of CNN, which subsequently serves as input for BiLSTM [39]. CNNs are utilized to extract spatial features from the dataset. Bidirectional Long Short-Term Memory (BiLSTM) networks are designed to process temporal or time-sequence information derived from the extracted features. The integration of these two models aims to enhance the performance of the hybrid model in managing complex data effectively. Figure 5 illustrates the network architecture of the CNN-BiLSTM hybrid model.

2.10. Model Evaluation

The evaluation of the model's performance was performed to determine the effectiveness of the classification model, employing metrics such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The data is structured in a confusion matrix, displaying the predicted results in conjunction with the actual classifications. The confusion matrix is classified into two types: the 2x2 binary

Confusion matrix, which facilitates comparison between predicted and actual classes, and the multiclass confusion matrix. The metrics employed in a multiclass confusion matrix consist of true positives (TP), false positives (FP), and false negatives (FN) [40]. The true positive (TP) value is obtained from the primary diagonal of the matrix. The false negative (FN) is derived from the corresponding row, while the false positive (FP) is obtained from the relevant column.

The classification model's performance is evaluated through the confusion matrix, which includes metrics like accuracy, precision, recall, and F1-score. Accuracy quantifies the proportion of correct predictions within a classification framework, whereas precision defines the ratio of true positive predictions to

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

the overall count of positive predictions generated. Recall is defined as the proportion of true positive identifications relative to the total number of actual positive instances. The F1-score represents the harmonic mean of precision and recall, frequently applied in scenarios characterized by unbalanced class distributions. The metrics serve as a standard for evaluating the performance of the developed model [41]. The metrics are expressed mathematically in Equations (5), (6), (7), and (8).

$$Accuracy = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i + TN_i} \quad (5)$$

$$Precision = \frac{TP_i}{TP_i + FP_i} \quad (6)$$

$$Recall = \frac{TP_i}{TP_i + FN_i} \quad (7)$$

$$F1\text{-}score = \frac{2 \times Recall_i \times Precision_i}{(Recall_i + Precision_i)} \quad (8)$$

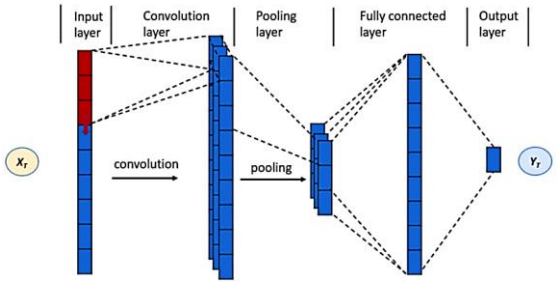

Figure 3. *CNN network architecture* [42]

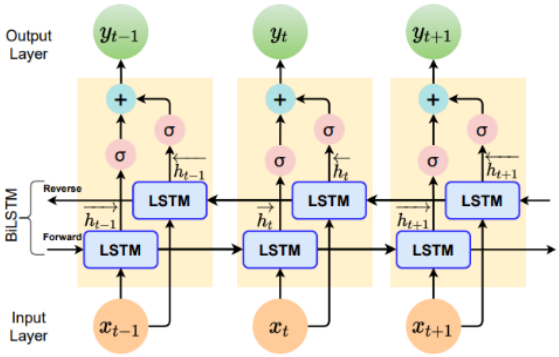

Figure 4. *BiLSTM network architecture* [42]

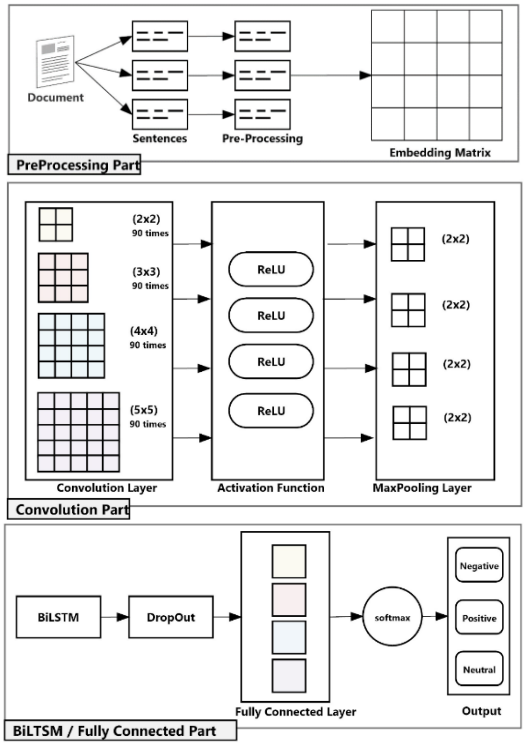

Figure 5. *CNN-BiLSTM network architecture* [23]

3. RESULTS AND DISCUSSION

3.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) represents the preliminary phase in data analysis, focusing on the comprehension of the dataset's characteristics and underlying patterns. This study conducted exploratory data analysis (EDA) to compare the proportions of ratings, aiming to evaluate the impact of the Halodoc application during and after the COVID-19 period. This method aims to collect data on changes in user sentiment concerning the application. Figure 6 illustrates the results of the visualization, which compares the proportion of ratings obtained from data collected during the COVID-19 period with those from the subsequent timeframe.

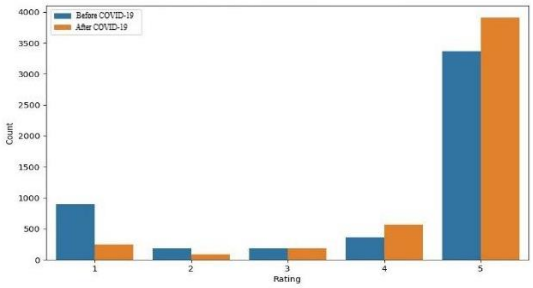

Figure 6. *Comparative analysis of rating proportions before and after the COVID-19 pandemic*

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

Figure 6 presents the changes in the percentage of ratings obtained from the data collected during and after the COVID-19 period. The calculation of percentage change for each assessment category is derived from the variation in the proportion of the assessment. Table 3 presents the percentage change of this value, reflecting both increases and decreases in the number of ratings.

Table 3. *Analysis of rating proportions pre and post-COVID-19*

Rating	Pre COVID-19	Post COVID-19	Percentage
1	251	897	257.37%
2	89	189	112.36%
3	184	187	1.63%
4	567	364	-35.80%
5	3909	3363	-13.97%

Analysis of Table 3 indicates that the Halodoc application exhibited increased user satisfaction during the COVID-19 period relative to the post-pandemic phase. The observed phenomenon may result from heightened demand for healthcare services during the COVID-19 period, characterized by restricted access to in-person healthcare facilities. This emergency condition requires heightened user engagement with applications such as Halodoc, thereby enhancing the overall user experience. Following the pandemic, there may be a reduction in the demand for healthcare services, potentially impacting user satisfaction levels with the application.

3.2. Data Pre-processing

The data pre-processing procedure is performed to ready the data for application using the deep learning method. The first step in this process is to remove duplicate data, as redundancy may lead to bias in the analysis. The next phase encompasses several essential processes, such as case folding for standardizing letter formatting, data cleaning to remove unnecessary characters or symbols, stopword removal to reduce the occurrence of ordinary words with little significance, and categorical coding to convert categorical variables into numerical format. Tokenization involves the segmentation of text into its fundamental units, known as tokens. Table 4 presents a comprehensive summary of the results derived from the pre-processing procedure.

Table 4. *Pre-processing results*

Before	After
saya sangat puas dengan aplikasi ini. dokternya ramah2 dan juga banyak banget artikel2 tentang kesehatan yg sangat bermanfaat. terimakasih halodoc,sudah memberikan yg terbaik untuk masyarakat 😁😀😄	*saya sangat puas dengan aplikasi ini dokternya ramah dan juga banyak banget artikel tentang kesehatan yg sangat bermanfaat terimakasih halodoc sudah memberikan yg terbaik untuk masyarakat*
Eng. ver: I am highly satisfied with this application. The doctors are exceptionally friendly, and there is an abundance of informative health articles available. Thank you, Halodoc, for providing excellent services to the community 😁😀😄	Eng. ver: i am highly satisfied with this application. the doctors are exceptionally friendly, and there is an abundance of informative health articles available. thank you, halodoc, for providing excellent services to the community

3.3. Data Resampling

Moreover, the study identified a data imbalance, as indicated by the disparity in the number of classes within the rating variables across the COVID-19 period and subsequent data collection intervals. In order to tackle this issue and enhance the precision of the categorization outcomes, the researcher employed data resampling techniques. In order to rectify this imbalance, the present study employed the Random Oversampling (ROS) methodology. This methodology involves the random replication of data from the minority class to achieve parity with the volume of data present in the majority class. The results of implementing random oversampling on data gathered during and subsequent to the COVID-19 pandemic are illustrated in Figures 7 and 8.

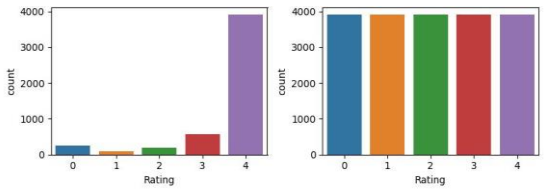

Figure 7. Review data during COVID-19 for (a) before resampling, (b) after resampling

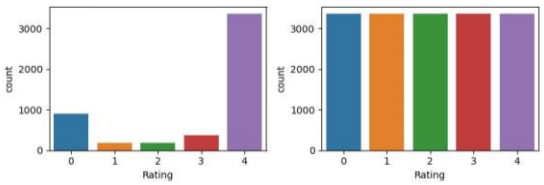

Figure 8. *Review data after COVID-19 for (a) before resampling, (b) after resampling*

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

3.4. Word Embedding

The data that has completed the pre-processing stage is subsequently tokenized, followed by the implementation of a word embedding process. The model employed for word embedding is Word2Vec. This model utilizes the Continuous Bag of Words (CBOW) architecture to generate vector representations of words by predicting target words based on their contextual environments. The variability in the length of user review texts requires the implementation of sequence padding procedures. This padding sequence involves the integration of a zero value into the vector representation of the text, thereby guaranteeing that all vectors uphold a uniform length. Table 5 presents an illustration of the outcomes derived from the word embedding process utilizing the Word2Vec model, following the padding sequence phase.

Table 5. *Results of word embeddings generated using the Word2Vec model*

Before	After
aplikasinya bagus dokternya ramah ... Eng. ver: the application is well designed the doctors are courteous	[0, 0, 0, ..., 16, 350, 14]
terimakasih aplikasi sangat membantu ... Eng. ver: thank you the application has proven to be helpful	[0, 0, 0, ..., 68, 0, 28]
sangat puas aplikasi dokternya ramah ... Eng. ver: i am highly satisfied with the application the doctors friendly	[0, 0, 0, ..., 64, 214, 198]

3.5. Data Splitting

After completing the word embedding phase, the next step is to facilitate data sharing. This division is critical for ensuring that the developed model can adapt and operate efficiently on data that has not been encountered before. The study's data was organized into three main categories: training data, validation data, and test data. The dataset is partitioned, with 90% allocated for training and the remaining 10% reserved for testing. The dataset is partitioned into two segments: 85% is assigned for model training, while the remaining 15% is reserved for model validation. Table 6 displays the outcomes of the data sharing, covering the COVID-19 period and the following timeframe.

Table 6. *Data splitting scheme*

Data	Data Splitting		Total
During COVID-19	Model 90%	Training 85%	12469
		Validation 15%	2201
	Testing 10%		1630
After COVID-19	Model 90%	Training 85%	14836
		Validation 15%	2619
	Testing 10%		1940

3.6. Hyperparameter Tuning

Hyperparameter tuning aims to identify the most practical combination of model parameters that produce the highest performance results. The grid search method is utilized in this process to train the model by assessing all potential combinations of specified hyperparameter values. The performance score for each resulting model is computed for every combination, facilitating the identification of the model that exhibits optimal results. The results of hyperparameter tuning evaluations performed on data collected both during and after the COVID-19 pandemic are presented in Table 7 and Table 8.

Table 7. *Outcomes of hyperparameter optimization for the COVID-19 dataset*

Batch size	Learning rate	Accuracy	Loss
32	1×10^{-2}	0,990279	0,008186
32	1×10^{-3}	0,994372	0,004221
64	1×10^{-2}	0,985930	0,007930
64	$\mathbf{1 \times 10^{-3}}$	**0,996419**	**0,000512**

Table 8. *Outcomes of hyperparameter optimization for the post-COVID-19 dataset*

Batch size	Learning rate	Accuracy	Loss
32	1×10^{-2}	0,985058	0,000669
32	1×10^{-3}	0,990113	0,002007
64	1×10^{-2}	0,990485	0,002528
64	$\mathbf{1 \times 10^{-3}}$	**0,994573**	**0,000669**

3.7. Hybrid CNN-BiLSTM Model

The identified optimal parameter values are employed to develop a sentiment analysis model. The model developed is a hybrid architecture that combines Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (BiLSTM), consisting of multiple layers. The architecture includes an embedding layer that converts words into numerical vectors, a convolution

Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (209-225)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.42762

layer that extracts essential features from the data, and a pooling layer that reduces the dimensions of the data. A BiLSTM layer is then implemented to capture the temporal relationships present in the data. A flattened layer is then applied to prepare the output for the fully connected layer. The output layer generates the final prediction. The graph presented below depicts the recorded loss value, validation loss, accuracy, and validation accuracy throughout the model training process.

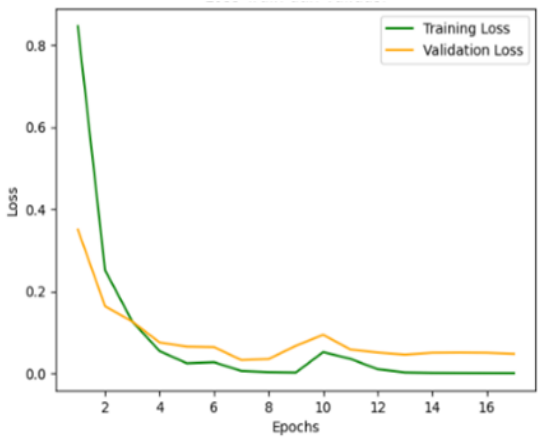

Figure 9. *Loss and validation of loss metrics during the COVID-19 period*

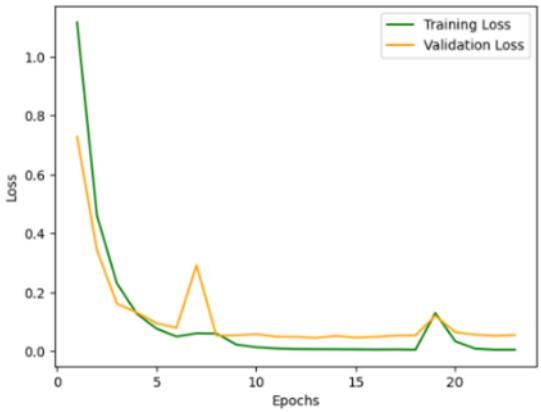

Figure 10. *Loss and validation of loss metrics after the COVID-19 period*

The results of model training on data during and after COVID-19 show a consistent pattern in evaluation metrics. Loss values and loss validation decreased significantly with each epoch, indicating that the model learned well from the data provided. In addition, the accuracy values and accuracy validation for both datasets also showed an upward trend that was in line with each epoch. That indicates that the developed CNN-BiLSTM hybrid model does not overfit and shows satisfactory performance. A multiclass confusion matrix was used to evaluate the performance of the model in more depth. The results of the confusion matrix for data during and after COVID-19 can be seen in Figures 11 and 12.

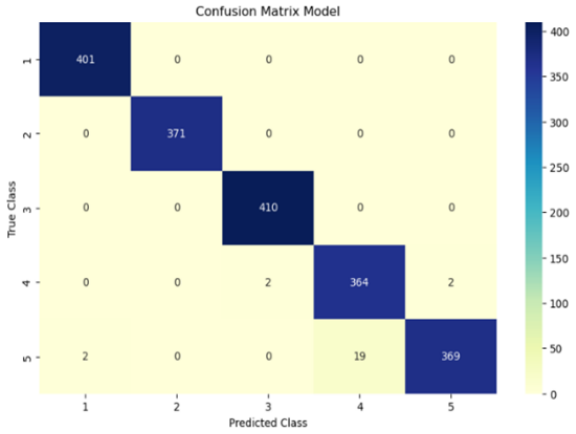

Figure 11. *Confusion matrix of the model during COVID-19*

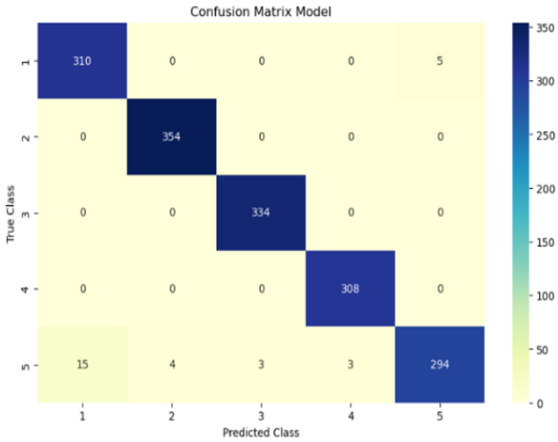

Figure 12. *Confusion matrix of the model after COVID-19*

The analysis performed using a confusion matrix reveals that the CNN-BiLSTM hybrid model exhibits enhanced performance in classifying ratings based on user reviews of the Halodoc application. The model achieved an accuracy of 98.71% on data collected during the COVID-19 pandemic, which later decreased to 98.16% on data gathered post-pandemic. However, misclassification analysis indicates that some instances were incorrectly classified, particularly in ratings 4 and 5. For instance, during the pandemic, there were cases where class 4 ratings were misclassified as class 3 or 5, and class 5 ratings were misclassified as class 1 or 4. Similarly, post-pandemic data also exhibited misclassifications, with class 1 being predicted as class 5 and class 5 being incorrectly classified as classes 1, 2, 3, or 4. Despite these misclassifications, the data demonstrates the model's capability to recognize patterns and

sentiments in user reviews, thus delivering an accurate depiction of user satisfaction with the application.

The evaluation of the model's performance demonstrated significant precision, recall, and F1-score metrics. The metrics provide a comprehensive evaluation of the model's ability to accurately and fairly classify both positive and negative instances. Tables 9 and 10 provide classification reports for data collected during and after the COVID-19 period, offering a comprehensive analysis of the model's performance for each timeframe. The CNN-BiLSTM hybrid model exhibited efficacy in understanding and analyzing user reviews during both periods.

Table 9. *Classification report model during COVID-19*

Rating	Precision	Recall	F1-score
1	0,98	1,00	0,99
2	1,00	1,00	1,00
3	0,99	1,00	1,00
4	0,93	0,96	0,95
5	0,96	0,91	0,93

Table 10. *Clasification report model after COVID-19*

Rating	Precision	Recall	F1-score
1	0,94	0,97	0,95
2	0,99	1,00	1,00
3	0,99	1,00	0,99
4	0,97	0,97	0,97
5	0,95	0,89	0,92

3.8. Comparison of Deep Learning Models and Real-World Challenges

The performance of the proposed Hybrid CNN-BiLSTM model is compared with other deep learning models, including MLP, CNN, LSTM, CNN-LSTM, and other hybrid models. The comparison, illustrated in Table 11, is based on accuracy across various sentiment analysis and classification datasets.

Traditional models such as MLP, CNN, and LSTM show varying accuracy levels, with Ali et al. (2019) reporting 86.74% for MLP, 87.70% for CNN, and 86.64% for LSTM on the IMDB dataset. Hybrid CNN-LSTM models improve performance, with Ali et al. (2019) achieving 89.20% and Jain et al. (2021) reaching 91.30% on different sentiment analysis tasks. More advanced hybrid CNN-BiLSTM models further enhance accuracy, as seen in Rhanoi et al. (2019) with 90.66%, Xiaoyan et al. (2022) with 95.65%, Singh et al. (2024), Susandri et al. (2024), and Ramaziyah & Setiawan (2024) with 71.60 - 99.00% accuracy in various domains.

The proposed Hybrid CNN-BiLSTM model demonstrates superior performance, achieving 98.16% and 98.71% accuracy in evaluating user satisfaction in the Halodoc application during pre- and post-pandemic COVID-19 periods. These results indicate its robustness in capturing user sentiments, highlighting its effectiveness in analyzing healthcare service feedback. Compared to previous hybrid models, the proposed method offers higher accuracy, making it a more reliable tool for assessing user satisfaction and enhancing digital healthcare services.

Despite its high accuracy, real-world deployment of this model in evaluating user satisfaction within the Halodoc application poses several challenges, as their decision-making processes are often difficult to explain to healthcare providers and stakeholders. Compliance with data privacy regulations is also crucial when processing sensitive user feedback. Addressing process challenges. Computational costs associated with deep learning models can limit large-scale implementation, particularly on resource-constrained systems. Additionally, the interpretability of deep learning models remains these challenges requires optimizing model efficiency and ensuring robust data security to enhance trust and usability in sentiment analysis for healthcare applications like Halodoc.

Table 11. *Comparison model with other deep learning models*

Model	Authors	Data	Accuracy (%)
MLP	Ali et al. (2019)	IMDB dataset for sentiment analysis	86.74
CNN			87.70
LSTM			86.64
Hybrid CNN-LSTM			89.20
Hybrid CNN-BiLSTM	Rhanoi et al. (2019)	French newspaper articles	90.66
Hybrid CNN-LSTM	Jain et al. (2021)	Airline Quality database and twitter for consumer sentiment	91.30
Hybrid CNN-BiLSTM	Xiaoyan et al. (2022)	Sourced from social media platforms for sentiment classification	95.65
Hybrid CNN-LSTM	Ashraf et al. (2023)	Music clips for music genres classification	87.20
Hybrid CNN-BiLSTM			88.00
Hybrid CNN-BiLSTM	Singh et al. (2024)	Emotion analysis in Urdu, one of the primary languages in Asia	99.00
CNN	Susandri et al. (2024)	Sentiment classification on WhatsApp group	75.00
Simple RNN			80.00
LSTM			81.00
BiLSTM			82.00
Hybrid CNN-BiLSTM			88.00
Hybrid BiLSTM - CNN	Ramaziyah & Setiawan (2024)	Source from Twitter for sentiment analysis	71.60
Hybrid CNN-BiLSTM			72.14
Hybrid CNN-BiLSTM	**Proposed Method**	**Source from Halodoc application during & post pandemic COVID-19**	**98.16 & 98.71**

CONCLUSION

The analysis and discussion indicate that the Halodoc application demonstrated heightened user satisfaction during the COVID-19 period when compared to the post-pandemic phase. This increase can be linked to the heightened needs of users during that specific timeframe. This study evaluates the effectiveness of a hybrid model integrating CNN and BiLSTM for classifying review ratings in the Halodoc application. The findings indicate that the CNN-BiLSTM hybrid model, following hyperparameter optimization, attained an accuracy of 98.71% for data collected during the COVID-19 period and 98.16% for data obtained subsequently. The optimal parameters identified include a batch size of 64 and a learning rate of 1x10-3. The loss graph demonstrates a steady decline in both loss value and validation loss across the epochs, signifying effective model performance in the classification of review ratings.

Future work should include the implementation of data resampling techniques, such as SMOTE or a hybrid method that integrates random oversampling and undersampling, to reduce the impact of outliers and tackle data imbalances. Furthermore, augmenting the quantity of hyperparameters assessed throughout the tuning process has the potential to improve the model's precision in determining the suitable parameters for classification.

REFERENCES

[1] R. V. Silalahi, N. Hartono, and M. A. Tumpak, "Profile and preferences users of doctors consultation application in Indonesia," in *IOP Conference Series: Earth and Environmental Science*, 2018. doi: 10.1088/1755-1315/195/1/012069.

[2] N. Hartono, L. Laurence, and T. O. Tedja, "Development initial model of intention to use Halodoc application using PLS-SEM," *Int. Conf. Informatics, Technol. Eng. 2019*, no. August, pp. 63–70, 2019.

[3] O. Thinnukool, P. Khuwuthyakorn, P. Wientong, and T. Panityakul, "Non-prescription medicine mobile healthcare application: Smartphone-based software design and development review," 2017. doi: 10.3991/ijim.v11i5.7123.

[4] A. J. Barton, "The regulation of mobile health applications," 2012. doi: 10.1186/1741-7015-10-46.

[5] B. Martínez-Pérez, I. De La Torre-Díez, and M. López-Coronado, "Mobile health applications for the most prevalent conditions by the world health organization: Review and analysis," 2013. doi: 10.2196/jmir.2600.

[6] H. Abaza and M. Marschollek, "mHealth application areas and technology combinations: A comparison of literature from high and low/middle income countries," 2017. doi: 10.3414/ME17-05-0003.

[7] M. A. Kushendriawan, H. B. Santoso, P. O. H. Putra, and M. Schrepp, "Evaluating User Experience of a Mobile Health Application 'Halodoc' using User Experience Questionnaire and Usability Testing," *J. Sist. Inf.*, vol. 17, no. 1, pp. 58–71, 2021, doi: 10.21609/jsi.v17i1.1063.

[8] M. Christian, E. Retno Indriyarti, S. Sunarno, and S. Wibowo, "Determinants of Satisfaction Using Healthcare Application: A Study on Young Halodoc Users in Jakarta During the COVID-19 Pandemic," vol. 2, no. 1, pp. 36–48, Aug. 2022, doi: 10.31098/quant.947.

[9] S. Chatterjee, "Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents," *Decis. Support Syst.*, vol. 119, pp. 14–22, 2019, doi: 10.1016/j.dss.2019.02.008.

[10] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decis. Anal. J.*, vol. 3, p. 100073, 2022, doi: 10.1016/j.dajour.2022.100073.

[11] J. Hirschberg and C. D. Manning, "Advances in natural language processing," 2015. doi: 10.1126/science.aaa8685.

[12] J. Berger, A. Humphreys, S. Ludwig, W. W. Moe, O. Netzer, and D. A. Schweidel, "Uniting the Tribes: Using Text for Marketing Insight," *J. Mark.*, vol. 84, no. 1, pp. 1–25, 2020, doi: 10.1177/0022242919873106.

[13] J. Wang *et al.*, "Global evidence of expressed sentiment alterations during the COVID-19 pandemic," *Nat. Hum. Behav.*, vol. 6, no. 3, pp. 349–358, 2022, doi: 10.1038/s41562-022-01312-y.

[14] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a Feeling: Accuracy and Application of Sentiment Analysis," *Int. J. Res. Mark.*, vol. 40, no. 1, pp. 75–87, 2023, doi: 10.1016/j.ijresmar.2022.05.005.

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

[16] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, "Fundamental concepts of convolutional neural network," in *Intelligent Systems Reference Library*, vol. 172, 2019, pp. 519–567. doi: 10.1007/978-3-030-32644-9_36.

[17] M. Vineethmohan, P. Hemanth, M. Mounica, and P. Lakshmi Prasanna, "Image classification using deep learning," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. 6, pp. 1–10, Mar. 2020, doi: 10.5373/JARDCS/V12I6/S20201001.

[18] H. Fan, M. Jiang, L. Xu, H. Zhu, J. Cheng, and J. Jiang, "Comparison of long short term memory networks and the hydrological model in runoff simulation," *Water (Switzerland)*, vol. 12, no. 1, 2020, doi:

10.3390/w12010175.

[19] S. M. Mousavi, M. Ghasemi, M. D. Manshadi, and A. Mosavi, "Deep learning for wave energy converter modeling using long short-term memory," *Mathematics*, vol. 9, no. 8, 2021, doi: 10.3390/math9080871.

[20] L. Shan, Y. Liu, M. Tang, M. Yang, and X. Bai, "CNN-BiLSTM hybrid neural networks with attention mechanism for well log prediction," *J. Pet. Sci. Eng.*, vol. 205, no. December 2020, p. 108838, 2021, doi: 10.1016/j.petrol.2021.108838.

[21] N. M. Ali, M. M. A. El Hamid, and A. Youssif, "Sentiment Analysis for Movies Reviews Dataset Using Deep Learning Models," *Int. J. Data Min. Knowl. Manag. Process*, vol. 09, no. 03, pp. 19–27, 2019, doi: 10.5121/ijdkp.2019.9302.

[22] P. K. Jain, V. Saravanan, and R. Pamula, "A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, 2021, doi: 10.1145/3457206.

[23] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM Model for Document-Level Sentiment Analysis," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 3, pp. 832–847, 2019, doi: 10.3390/make1030048.

[24] L. Xiaoyan, R. C. Raga, and S. Xuemei, "GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews," *J. Sensors*, vol. 2022, 2022, doi: 10.1155/2022/7212366.

[25] M. Ashraf *et al.*, "A Hybrid CNN and RNN Variant Model for Music Classification," *Appl. Sci.*, vol. 13, no. 3, 2023, doi: 10.3390/app13031476.

[26] N. Singh, U. C. Jaiswal, and R. Singh, "Detecting Sarcasm Text in Sentiment Analysis Using Hybrid Machine Learning Approach," *Int. J. Intell. Syst. Appl.*, vol. 16, no. 4, pp. 72–85, 2024, doi: 10.5815/ijisa.2024.04.05.

[27] Y. A. Ramaziyah and E. B. Setiawan, "Hybrid Deep Learning CNN and BiLSTM with FastText as Feature Expansion for Sentiment Analysis in President Election 2024," in *COMNETSAT 2024 - IEEE International Conference on Communication, Networks and Satellite*, 2024, pp. 176–183. doi: 10.1109/COMNETSAT63286.2024.10862946.

[28] S. Susandri, S. Defit, and M. Tajuddin, "Enhancing Text Sentiment Classification with Hybrid CNN-BiLSTM Model on WhatsApp Group," *J. Adv. Inf. Technol.*, vol. 15, no. 3, pp. 355–363, 2024, doi: 10.12720/jait.15.3.355-363.

[29] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2vec model analysis for semantic similarities in English words," in *Procedia Computer Science*, 2019, pp. 160–167. doi: 10.1016/j.procs.2019.08.153.

[30] I. Surjandari, C. Megawati, A. Dhini, and I. B. N. Sanditya Hardaya, "Application of text mining for classification of textual reports: A study of Indonesia's national complaint handling system," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*, 2016, pp. 1147–1156.

[31] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," 2021. doi: 10.1007/s42979-021-00592-x.

[32] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.

[33] S. Sharma, S. Sharma, and A. Athaiya, "Activation Functions in Neural Networks," *Int. J. Eng. Appl. Sci. Technol.*, vol. 04, no. 12, pp. 310–316, 2020, doi: 10.33564/ijeast.2020.v04i12.054.

[34] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol. 8, no. 4, 2021, doi: 10.3390/informatics8040079.

[35] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*, 4th ed., no.

August. 2011. doi: 10.1016/C2009-0-61819-5.

[36] D. Gao, X. Liu, Z. Zhu, and Q. Yang, "A hybrid CNN-BiLSTM approach for remaining useful life prediction of EVs lithium-Ion battery," *Meas. Control (United Kingdom)*, 2022, doi: 10.1177/00202940221103622.

[37] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 2021, doi: 10.1016/j.neucom.2020.04.157.

[38] A. F. Hidayatullah, S. Cahyaningtyas, and R. D. Pamungkas, "Attention-based CNN-BiLSTM for Dialect Identification on Javanese Text," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, pp. 317–324, 2020, doi: 10.22219/kinetik.v5i4.1121.

[39] R. Bharal and O. V. V. Krishna, "Social Media Sentiment Analysis Using CNN-BiLSTM," *Int. J. Sci. Res.*, vol. 10, no. 9, pp. 656–661, 2020, doi: 10.21275/SR21913110537.

[40] D. H. N. Aini, D. Kurniasari, A. Nuryaman, and M. Usman, "Implementation of Artificial Neural Network With Backpropagation Algorithm for Rating Classification on Sales of Blackmores in Tokopedia," *J. Tek. Inform.*, vol. 4, no. 2, pp. 365–372, 2023, doi: 10.52436/1.jutif.2023.4.2.539.

[41] M. Hossin and M. . Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[42] B. Bohara, R. I. Fernandez, V. Gollapudi, and X. Li, "Short-Term Aggregated Residential Load Forecasting using BiLSTM and CNN-BiLSTM," *2022 Int. Conf. Innov. Intell. Informatics, Comput. Technol. 3ICT 2022*, pp. 37–43, 2022, doi: 10.1109/3ICT56508.2022.9990696.