Enhancing Text Sentiment Classification with Hybrid CNN-BiLSTM Model on WhatsApp Group

Susandri Susandri [1,*], Sarjon Defit [2], and Muhammad Tajuddin [3]

[1] Faculty of Computer Science, STMIK Amik Riau, Pekanbaru, Indonesia
[2] Faculty of Computer Science, UPI YPTK Padang, Padang, Indonesia
[3] Faculty of Computer Science, Bumigora University, NTB, Indonesia
Email: susandri@sar.ac.id (S.S.); sarjon_defit@upiyptk.ac.id (S.D.); tajuddin@universitasbumigora.ac.id (M.T.)
*Corresponding author

***Abstract*—Large amounts of data are generated from social media. The need to extract meaningful information from big data, classify it into different categories, and predict user sentiment is crucial. Text classification is a representative research topic in the field of natural language processing that categorizes unstructured text data into sentiments to make it more meaningful. Improving word and text category accuracy requires more precise text classification methods. Deep Learning models developed and implemented in this field have shown progress, but further improvement is still needed. This paper utilizes the NLP process on a WhatsApp group dataset to determine sentiment, testing it with five Deep Learning models: Neural Network, Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory, Convolutional Neural Network (CNN), and proposes a hybrid CNN-BiLSTM model. The proposed model employs feature extraction and a hybrid architecture with activations, dropouts, filters, kernel sizes, and different layers to classify text sentiment. To verify the performance of the proposed model, it is compared with previous studies. In single-model testing, the Long Short-Term Memory and BiLSTM achieves the best accuracy of 81%. Meanwhile, the proposed model has reached an accuracy of 88% on the utilized dataset. By comparing the performance of the proposed model with previous studies, the proposed model offers better sentiment classification performance.**

***Keywords*—text classification, WhatsApp group, hybrid CNN-BiLSTM**

I. Introduction

Text Classification (TC) refers to the systematic categorization of textual data into distinct groups based on their inherent properties and characteristics. Through automated analysis, TC effectively examined text and reliably assigned pre-established categories. This procedure is of utmost importance for facilitating the processing and extraction of valuable information from raw and unstructured textual data [1, 2]. TC encompasses three distinct system types: rule-based, machine-learning-based, and hybrid approaches [3, 4]. Rule-based systems employ a collection of predefined rules to effectively categorize text into organized groups. On the other hand, machine learning-based systems rely on prior observations and patterns to classify text. Hybrid systems, as the name implies, combine elements of rule-based and machine-learning-based systems, utilizing both trained classifiers to improve classification accuracy and efficacy.

Improving word and text category accuracy requires more accurate text classification methods [3]. In the realm of text classification, the application and advancement of Deep Learning (DL) models have emerged [5, 6]. DL models commonly employed for text classification include Rule-embedded Neural Networks (ReNNs), Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). These models have demonstrated superiority and effectiveness in accurately categorizing textual data. The use of Word2Vec and GloVe has also improved classification accuracy at the sentence, paragraph, and document levels [6, 7].

Many researchers have dedicated their efforts to enhancing the efficacy of Natural Language Processing (NLP) tasks through the development of Deep Learning (DL) frameworks [7, 8]. Notably, CNN, Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM) techniques attained an accuracy rate of 77.4%, with CNN exhibiting the most favorable average performance [9]. These advancements in DL frameworks have contributed significantly to the improved performance and accuracy of NLP tasks in various applications. Researchers continue to develop hybrid DL models to achieve better results.

The hybrid Bayesian Network and RNN (BN-RNN) method achieved evaluation metrics ranging from 73.4% to 78.4% [10]. The combination of Long Short-Term Memory and Neural Networks (LSTM-NN) achieved an accuracy of 66.5%, precision ranging from 65.9% to 83.7%, recall ranging from 79.9% to 81.1%, and F1-Score ranging from 71.6% to 80.0% [11]. In contrast, the combined Bi-LSTM and MLP (BiLSTM-MLP) approach demonstrated a commendable accuracy of 88.3% and an

Manuscript received September 25, 2023; revised November 1, 2023; accepted November 16, 2023; published March 14, 2024.

impressive F1-Score of 85.8% [12]. This achievement highlights the efficacy and robustness of the BiLSTM-MLP method in accurately categorizing text data, demonstrating its potential for enhancing the performance of text classification tasks.

Researchers have also developed hybrid DL models. In comparison to existing techniques, the MTL-MSCNN-LSTM model, a hybrid multi-scale CNN, and LSTM approach for multitask multi-scale sentiment classification demonstrated superior performance with accuracy rates ranging from 82.88% to 88.50% [13, 14]. This signifies the effectiveness and superiority of the MTL-MSCNN-LSTM model in accurately classifying sentiments across multiple tasks and varying scales, showcasing its potential for advancing sentiment classification methodologies. This notion was substantiated by Ayo *et al*. [15] in their comprehensive analysis of hybrid approaches, wherein they extensively evaluated these methods using key performance metrics such as Accuracy, Precision, Recall, and F1-Score. Their examination further reinforces the effectiveness and validity of hybrid methodologies in addressing various challenges and yielding favorable outcomes in diverse applications.

Combining CNN and BiLSTM in one architecture enables more comprehensive and in-depth data processing. CNN helps extract spatial features from the data, while BiLSTM helps extract temporal or contextual features [16]. Researchers have implemented hybrid CNN-BiLSTM models [17, 18], and using Word2Vec, GloVe, and fasttext. They used an embedding size of 300, Softmax activation, dropout, and recurrent dropout rates of 0.5 and 0.4, an SGD optimizer, 50 epochs, a filter size of 512, a kernel size of 3, embedding, CNN max pooling, BiLSTM, and dense layers, resulting in an accuracy of 82% [13].

Although the integration of DL hybrids by researchers has shown progress, there are still limitations, such as handling long input and unnecessary convolution operations for NLP [8], as well as data loss with increasing data size, requiring more training data and time [10].

This research introduces a novel approach of employing a Hybrid CNN-BiLSTM model, which integrates distinct feature extraction techniques and varying kernel sizes, for the purpose of analyzing the Whatsapp Group (WAG) dataset. By blending the capabilities of CNN and BiLSTM, this model aims to enhance the accuracy and effectiveness of data analysis and classification tasks on a specific dataset under investigation. This model leverages the advantages of both CNN for extracting local features and BiLSTM for handling long-term dependencies. Unlike the basic LSTM that only uses past information, when dealing with sequential time-series data, the BiLSTM model integrates future and past information, thereby incorporating additional context and enhancing the prediction accuracy.

Furthermore, the model incorporates convolutional kernels of various sizes to effectively capture local temporal dependencies within sequential data. To assess the performance of the model, TC experiments utilized a confusion matrix. Given the slight disparities in data format compared with other social media data, the data require specific treatment. This study includes various treatments such as preprocessing, sentiment labeling, and feature extraction. This study makes several key contributions: (1) the development of a novel hybrid DL model for sentiment analysis, (2) evaluating the model's efficacy on WAG social media text, and (3) conducting performance comparisons between the proposed model and other DL models utilized for text classification.

The subsequent section of this article is structured as follows: Section II provides a comprehensive explanation of the proposed material and methods, encompassing data preprocessing, the feature extraction process, and the architecture of the proposed model. The outcomes of the experiments and the specific dataset employed in the proposed work are presented and discussed in Section III. Finally, Section IV concludes the study and offers insights into potential avenues for future research.

II. Literature Review

Previous studies related to text classification and sentiment analysis have been undertaken, and they adhere to the process of classifying text sentiment. These studies include:

- This investigation focused on Thai language sentiment analysis using a DL classification model [9]. The study aimed to improve Thai language sentiment analysis by comparing word embedding, POS tagging, and sentic features. Sentic features are essential for enhancing sentiment analysis, providing information about word polarity and subjectivity, thus improving sentiment context comprehension. Three distinct models were explored, with the LSTM model achieving the highest F1-Score (0.726) using POS-tag embedding and sentic features. The BiLSTM model, with word embedding, POS-tag one-hot encoding, and selected sentic features, achieved an F1-Score of 74.7%, while the CNN model, using word embedding, POS-tag one-hot encoding, and sentic features, performed best with an F1-Score of 81.7%. Further enhancements may be possible with additional deep learning features or models in future research.
- This study aimed to enhance text classification accuracy using a Bi-LSTM model that combines Word2Vec, CNN, and attention mechanisms [13]. The approach combined multiple effective text classification techniques: Word2Vec for richer word vectors, CNN for local feature extraction, and attention mechanisms for weight assignment to significant sections of the text. The model used parameters such as a skipgram embedding size of 300, a 0.2 dropout rate, batch size of 128, and the Adam optimizer. It achieved an average accuracy of 87.4%, with the highest F1-Score (90.1%) observed on a 13k-instance dataset. This study demonstrated the synergistic integration of these models enhances text classification accuracy. Note that having access to more extensive training data and further refinement time benefits the model.

- In a recent study, Salur and Aydin [17] introduced an innovative hybrid deep-learning model for sentiment classification. This model addresses dataset complexity and imbalances by synergistically combining CNN for local textual feature extraction and LSTM-based RNN for long-term contextual information. To enhance feature representations, the model includes character embedding and pre-trained embedding (Word2Vec, GloVe, FastText). Using tweets related to a Turkish GSM operator, the results showed that a single BiLSTM model achieved 80.44% accuracy with FastText embedding. However, the hybrid CNN-BiLSTM model, which incorporates character embedding and FastText, outperformed it with 82.14% accuracy. These results emphasize the superiority of the proposed hybrid model for sentiment classification, although it is specifically tailored to Turkish, Arabic, and Lithuanian languages.
- In a study by Naqvi *et al*. [18], sentiment analysis in Urdu text was addressed using deep learning. Their approach utilizes deep learning to enhance Urdu sentiment analysis accuracy, combining CNN for local feature extraction and RNN with LSTM for long-term context understanding. Four embeddings, Samar, CoNLL, pretrained, and self-trained Fast Text, were evaluated. The BiLSTM ATT model achieved the highest accuracy at 77.9%, while LSTM reached the highest precision (85.16%) with Samar embedding. These results highlight the effectiveness of their deep learning methods in Urdu sentiment analysis. Further advancement potential exists through exploring alternative methods. This section reinforces the background by providing evidence and discussing trends in the field. It should encompass all relevant studies and supporting evidence.

A hybrid CNN-BiLSTM model was developed using different feature extraction techniques and kernels for the WAG dataset. This model is an innovative approach because it combines the advantages of both the techniques. The CNN is effective in its ability to extract local features from data, whereas the BiLSTM model excels in handling long-term dependencies and leveraging both past and future information to improve prediction accuracy. To accommodate varying data lengths, the inclusion of padding techniques adds value to the model's capability. The incorporation of the Rectified Linear Unit (ReLU) and Softmax activation functions further enhanced the flexibility of data processing. With an embedding size of 300, the model could extract more intricate and complex features from the input data.

This unique and innovative approach for analyzing sequential data within the WAG dataset involves the utilization of four LSTM layers and the selection of 64 units in the final layer. Additionally, the model incorporates convolutional kernels of different sizes to capture local dependencies in sequential data. The novelty of this research lies in the development of a Hybrid CNN-BiLSTM model that combines various feature extraction techniques and incorporates convolutional kernels to effectively enhance the performance in the analysis of sequential data.

III. Materials and Methods

This study aims to enhance the performance of the hybrid CNN BiLSTM model, specifically for sentiment analysis in the TC domain. We evaluated this hybrid model thoroughly by considering the advantages and potentials of both the CNN and BiLSTM architectures. The methodology we adopted for this study focuses on crucial aspects, such as selecting appropriate data, performing meticulous labeling, developing robust feature vectors, and formulating the hybrid CNN BiLSTM approach. All these aspects contribute to a more precise and accurate sentiment analysis solution. We applied the proposed model to effectively predict the sentiment polarity of textual data and subsequently classify it based on the determined polarity

A. Dataset

We obtained the research dataset from the "Forum DTC Riau" WhatsApp group, a community comprising 134 owners of Daihatsu Taruna cars in the Riau region. We chose to use the "Forum DTC Riau" dataset (DTC) because it aligns with our research objectives and specific geographical characteristics. This dataset helps us understand unique preferences, issues, and perspectives within the region, which could potentially influence user sentiments and viewpoints. For comparison, we also tested the model with the Amazon Product Summaries dataset from Kaggle (PS), which consists of ten thousand data record. We followed similar preprocessing and model testing procedures for this dataset

This WAG was established on 10/11/2018, and the conversation data collected for analysis spanned from 16/3/2023 to 15/3/2023. The data extraction process involved using an OPPO A15 smartphone equipped with an Android Version 10 operating system. The smartphone employed an octa-core processor with 3 Gigabyte RAM capacity to facilitate the data extraction procedure. All conversation data from the WAG were exported in text file (txt) format and subsequently transmitted directly via email.

The text file containing the WhatsApp group data is unstructured data. It contains information about encrypted messages and calls exchanged between the group members, as well as the creation time of the group. This data needs to be readable and understandable by machines (computers), for which NLP is used to comprehend, classify, and extract opinions, sentiments, and emotions from natural language texts or data [19]. The NLP performed in this dissertation consists of two stages: data labelling and text preprocessing

In this study, the framework and methodology include several main stages: labeling, preprocessing, feature extraction, and the use of the proposed model. The labeling stage involves assigning labels or classifications to the data used, which facilitates the machine-learning process. Next,

a preprocessing stage was conducted to clean and prepare the data for further analysis. Subsequently, the feature extraction stage is performed to identify and extract important features from the data, which can aid in modeling and further analysis. Finally, the use of the proposed model involves implementation and testing of the developed model with the objective of achieving the desired results. By following these stages, this study aims to produce a comprehensive and accurate analysis based on carefully prepared and validated data

B. Labeling

Unstructured data is contained in the text file data from the WAG [20]. It also includes the WAG's creation time and details on the encrypted calls and messages that were made and received by members of the WAG. Dates, timestamps, cellphone numbers, names of WAG members, emojis (emoticons), and <Media omitted>, which reflects the traces of WAG members sending messages in the form of photographs or multimedia, are all included in the data. The message data doesn't have sentiment classification, so an emotion-based sentiment analysis method is used to determine the classification, using the SentiWordNet emotional lexicon [21, 22]. The following processes are used in the labeling sentiment process to build models and analyze emotional words and sentences:

- The first stage in data processing is to read and organize the data with regular expressions. Each line was read and broken into groups using commas as delimiters. A split function was used to retrieve the first item from each group. Following that, the aggregated data were subjected to further processing by tokenizing the date and time, author, and message tokens. The message itself consists of text, emoticons, and URL links sent by the group's members. These components are extracted and merged for further categorization into various categories: "Message", "emoji", and "urlcount". To do this, each line of data is separated into commas (,), hyphens (-), colons (:), and spaces (), allowing the necessary tokens to be retrieved and effectively stored within a data frame. The final outcomes of this process were presented through the utilization of the panda's library
- Labeling data entails classifying information into positive, neutral, and negative categories, with a focus on the message column. To initiate this labeling process, words within the messages are tokenized, and stop words are filtered out using the Natural Language Toolkit (NLTK) and Sastrawi libraries. Extraction was accomplished by adding a new column named Message_English, and the content of the message column was translated into English using the Google TransTranslator package.
- To facilitate the translation process, data were split into 250 rows. Using the nltk.sentiment.vader module, the resultant Message_English column was utilized for further extraction, allowing the insertion of positive, negative, neutral, and compound columns. Furthermore, the compound column is re-extracted to include a sentiment column, using the following guidelines: if the compound value is greater than or equal to 0.05, the sentiment is labeled as 'Positive'; if the compound value is less than or equal to –0.05, the sentiment is labeled as Ngative; and for all other values, the sentiment is labeled as Neutral. Finally, the labeled data were saved in a file named "DTCRiau_sentimen.csv" for future reference.
- The labeling for the PS dataset was conducted by extracting the Score column, with the stipulation that if the score value was >3, it was assigned a positive label; if the score =3, it was labeled as neutral, and if the score was <3, it was labeled as negative.

C. Preprocessing

Text preprocessing is a vital component of text classification, encompassing a range of techniques to prepare and transform text data for analysis. This process holds significant importance across diverse domains and languages [23]. Human-transferred text data needs to be in a machine-readable format for further analysis [24]. Preprocessing is done to eliminate issues that may interfere with the results in the subsequent data processing steps. The data used in the classification process is not always in an ideal condition. Preprocessing is necessary for labeled data before proceeding to the next stages.

The present study encompasses several essential pre-processing steps to enhance the quality and suitability of the data. These steps involve the removal of HTML codes and URLs from text messages to ensure cleaner data. Additionally, negation words were replaced with their respective antonyms (negation) to account for nuanced sentiment expressions. Neutral sentiment data were excluded from the dataset to focus solely on polarized sentiments. Moreover, punctuation marks were eliminated to minimize their potential influence on the subsequent analyses. Finally, lemmatization is performed to replace words with their base forms to improve consistency and linguistic accuracy.

D. Feature Extraction

In the field of NLP, a significant hurdle is developing a model capable of understanding the hierarchical representation of sentences in text data. This challenge arises primarily in the context of classification tasks and the extraction of relevant features [25]. Feature extraction involves capturing the characteristic features or attributes of a particular shape and then carrying out an analysis of the captured feature values. The process of feature extraction involves reducing the dimensionality of the output data to make it more manageable and enable efficient processing. By carefully selecting and/or combining variables into characteristics, the amount of data to be processed is effectively reduced.

Despite this reduction, the resulting features accurately and comprehensively characterize the underlying data. The data is organized into more manageable clusters. In this study, the following feature extraction processes were performed: Encoding is used to turn nominal or categorical

data into numerical data by dividing the data into training and testing sets, tokenization to break down sentences into smaller pieces of words or tokens, and padding to adjust the output dimensions to match the input dimensions.

E. Proposed Model

The dataset, which has been extracted and labeled with sentiment, was tested using single algorithms such as Neural Network, RNN, LSTM, BiLSTM, and CNN. Before the testing phase began, a series of pre-processing steps were carried out to ensure the cleanliness and suitability of the data. These steps included eliminating empty entries, URLs, emojis and punctuation. Once the data was effectively cleaned, feature extraction began, which included tokenization and encoding techniques.

The data were then divided into specific subsets through data partitioning and subjected to filling procedures to standardize the sequence lengths for further analysis and modeling. The test results were compared with the developed hybrid model.

The development of the hybrid CNN-BiLSTM model commenced with the labeling phase, pre-processing, and feature extraction, as previously described. We employed an embedding approach to represent text as low-dimensional numeric vectors with a padding size set at 300, which differs from previous approaches (see Table I).

TABLE I. HYPERPARAMETER OF THE PROPOSED HYBRID MODEL "SEQUENTIAL"

Layer (type)	Output Shape	Param #
Embedding	(None, 100, 300)	7,500,000
Conv1D	(None, 98, 200)	180,200
Bidirectional	(None, 98, 128)	135,680
Dropout	(None, 98, 128)	0
Bidirectional	(None, 128)	98,816
Dense	(None, 50)	6,450
Dense	(None, 50)	2,550
Flatten	(None, 50)	0
Dense	(None, 100)	5,100
Dense	(None, 2)	202

We compared this model's architecture with prior research, as documented in Table II. In each study, the model architecture was tailored to specific data characteristics. Jang *et al*. [13] utilized a hybrid CNN BiLSTM with Word2vec Skip Gram to process reviews of clothing and camera products. Salur and Aydin [17] employed tweets from Turkish GSM operator users using a model comprising CNN BiLSTM and character + fast text. Soumya and Pramod [26] explored sentiment in Malayalam Tweets using a model consisting of CNN BiLSTM and CNN LSTM

This hybrid model combines CNN layers for text feature extraction with BiLSTM layers to comprehend word context in the text. CNN, equipped with filters and kernels, identifies essential patterns within the text, whereas BiLSTM operates bidirectionally to understand word relationships. Non-linearity and the prevention of overfitting are introduced through the use of activation functions, specifically ReLU, and the incorporation of dropout mechanisms that randomly deactivate neurons during training.

We mitigated overfitting by adding a dropout layer following the initial BiLSTM layer. The dropout function introduces random neuron deactivation during training, and with a dropout rate set at 0, it ensures that active neurons prevent the model from closely fitting the training data, ultimately enhancing its generalization. Additionally, adding a dense layer with fewer units when paired with BiLSTM reduces the model's capacity to further prevent overfitting.

The combination of CNN and BiLSTM within a single architecture enables more comprehensive and in-depth processing of the data. CNN helps in extracting spatial features while BiLSTM helps in extracting temporal or contextual features [16]. The hybrid CNN-BiLSTM model begins with the labeling, preprocessing and feature extraction phases as explained previously.

The model is designed with embedding to represent the text as low-dimensional numerical vectors using padding with a size of 300. This embedding differs from the approaches used by [9, 13, 17] as shown in Table II.

TABLE II. PERFORMANCE OF THE EXISTING DL MODEL

Model	Embedding	F1-Score	Accuracy
CNN LSTM CNN Bi LSTM [9]	POS Tagging, sentic vevtor	81.7%	77.4%
CNN BiLSTM [13]	Word2vec Skip Gram	88.0%	87.6%
CNN BiLSTM [17]	Carakter +Fastext	89.0%	82.1%
CNN BiLSTM, CNN LSTM [26]	Sentiment Tagging, Wordvector	75.0%	85.5%

IV. RESULT AND DISCUSSION

The labeling of data utilizing the vader Sentiment library yielded sentiments categorized as positive, neutral, and negative. However, for the purpose of this research, the focus was specifically on positive and negative sentiments. The initially available data contained a total of 5,237 rows. To filter out the neutral sentiment labels, the dataset was reduced to 1,089 rows. This filtered dataset was then further divided based on the removal of stopwords.

The positive and negative sentiment data were then labeled using an encoder, resulting in subsequent columns representing these sentiments. The data was divided into training and testing sets for later analysis, with the training set including 80% of the data and the testing set containing the remaining 20% of the data. A comprehensive breakdown of the distribution of the training and testing data, with random_state = 69, is presented in Table III.

TABLE III. SPLITTING OF TRAINING DATASET AND TESTING DATASET

Feature	Training Dataset		Testing Dataset	
	DTC	PS	DTC	PS
Length	1,552	7,311	389	1,828
Dentences shape	1,552.0	7,311.0	389.0	1,828.0
Labels shape	1,552.2	7,311.2	389.2	1,828.2

The activation functions used are Softmax and ReLU. Dropout is set to 0.5 to reduce overfitting. The optimizer used is Adam, and the model is trained for 20 epochs. The model is made up of an embedding layer, a CNN layer, a BiLSTM layer, and four dense layers, as shown in Table I. The methodology used in this model distinguishes it from the approaches of previous studies [13, 17, 18]. The proposed model uses different filter sizes, this enables it to capture multiple local dependencies and so enhance the feature extraction procedure. Fig. 1 shows the architecture of the developed hybrid CNN-BiLSTM model. Comprehensive evaluations of both the single-DL and hybrid-DL models are carried out to assess their performance.

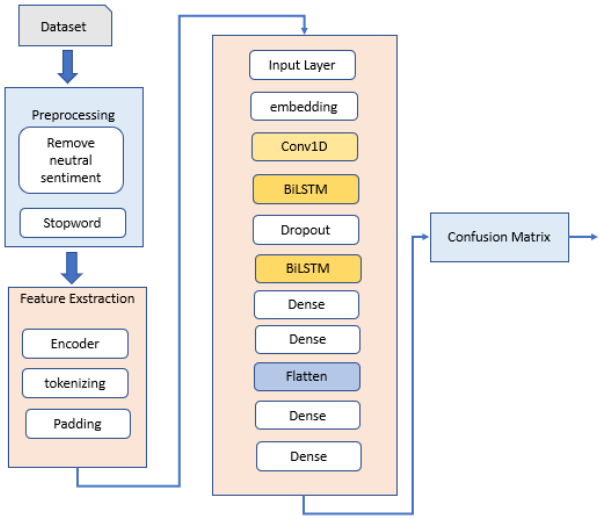

Fig. 1. The proposed architecture of the hybrid model.

The categorization results are given in the form of a confusion matrix. The model evaluation also gives accuracy, precision, recall, and F1-Score. These results are compared to previous research using the same method to produce better results using slightly different methods or phases. The following performance indicators are used in this work to assess the efficacy of the proposed single-DL and hybrid CNN BiLSTM models:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where, TP = True Positives, FN = False Negatives, TN = True Negatives, and FP = False Positives.

Precision: The proportion of real positive samples that were accurately recognized as positive out of the total number of positive samples.

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

The fraction of accurately recognized positive samples among all positive samples.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

F1-Score: A complete measure of the model's accuracy that may be determined as a harmonic average of precision and recall.

$$\text{F1-Score}=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

The PS dataset had a larger sample size than the DTC dataset in both the training and testing sets, indicating a stronger representation. Conversely, the DTC dataset had a smaller sample size, allowing for faster training, but with limitations in generalizing complex patterns (see Table III). The dataset division results revealed differences in sentiment distribution between the DTC and PS datasets. The PS dataset contained more positive sentiments, whereas the DTC dataset had a greater number of negative sentiments (see Fig. 2).

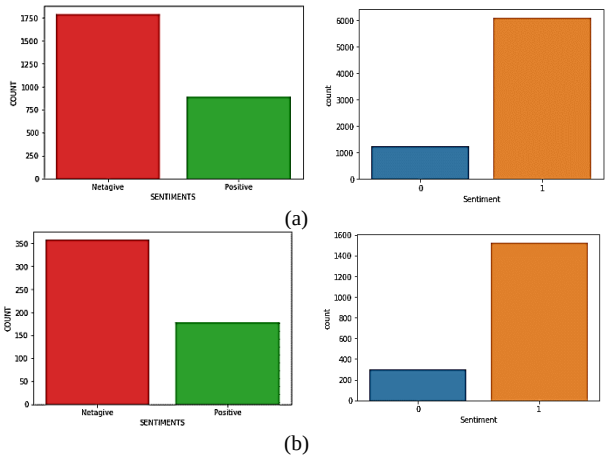

Fig. 2. Sentiment distribution of DTC and PS datasets; (a) Training dataset (b) Testing dataset.

Following the selection of the designated testing and training dataset, further preprocessing steps were employed to facilitate sub sequent analysis. The processing involved tokenization and data padding utilizing specific features, namely vocab_size = 25000, embedding_dim = 300, max_length = 100, trunc_type=post, and oov_tok = "<OOV>". These features were utilized to ensure efficient representation and standardization of the data for subsequent modeling and analysis purposes.

The tokenization and padding processes equalized the length of each sequence in the dataset. DTC has 871 training samples and 218 testing samples, whereas PS is significantly larger, with 7,311 training samples and 1,828 testing samples, with both datasets having a padding of size 100. The larger dataset size in PS has the potential for better generalization of more complex patterns, whereas DTC may have limitations owing to its smaller size.

For the evaluation of the extracted dataset with sentiment labels, various single deep learning classification models were employed. These models included Neural Network, RNN, LSTM, Bidirectional LSTM, and CNN. Modifications were made to the Preprocessing and Feature Extraction stages in order to assess the effectiveness of these models in handling the

dataset. By exploring these different models and adjusting these preprocessing and feature extraction techniques, a comprehensive evaluation of their performance was achieved. The Preprocessing stage involved changes in the order of Cleaning, Remove Neutral Sentiment, and Stop word.

The Feature Extraction stage involves processes such as label encoding, data splitting, tokenization, and addition. The deep learning models were tested and evaluated using an 80:20 data-split. Among the five classification models, the BiLTSM model achieved the highest accuracy of 0.81% on the DTC dataset and even higher accuracy on the PS dataset, as shown in Table IV, compared to other single deep learning models, with a dataset of 871 training data and 218 testing data from the "Forum DTC Riau" WhatsApp group.

TABLE IV. RESULTS DETAILS OF THE SINGLE DL MODEL AND THE PROPOSED MODEL

Dataset	Model	Accuracy	Precision	Recall	F1-Score
DTC	CNN	0.7%	0.68%	0.68%	0.68%
	Simple RNN	0.75%	0.77%	0.73%	0.74%
	LSTM	0.81%	0.83%	0.80%	0.81%
	BiLTSM	0.81%	0.85%	0.79%	0.80%
	Proposed Model	**0.8%**	**0.76%**	**0.82%**	**0.79%**
PS	CNN	0.75%	0.33%	0.25%	0.29%
	Simple RNN	0.80%	0.49%	0.59%	0.51%
	LSTM	0.81%	0.47%	0.50%	0.47%
	BiLTSM	0.82%	0.53%	0.49%	0.51%
	Proposed Model	**0.88%**	**0.76%**	**0.79%**	**0.78%**

This test can be further enhanced by using the proposed hybrid model. The proposed hybrid model achieved the highest accuracy among all the models, with a value of 0.88 on both datasets. However, its precision value is slightly lower at 0.76 compared to the other LSTM and BiLSTM models. In the DTC dataset, its precision value is lower (0.76) compared to LSTM and BiLSTM, while the PS dataset produced significant results with high accuracy and good recall and F1-Scores of approximately 0.79% and 0.78%, respectively. The proposed model excels in accuracy on both DTC and PS datasets, indicating the capability of the model to effectively classify the data.

The Feature Extraction stage involved processes such as Encoder Labeling, Data Splitting, Tokenization, and Padding. The DL models were tested and evaluated using an 80:20 data split. Among the five classification models, CNN, BiLTSM model achieved the highest accuracy of 0.81% as shown in Table IV compared to other single DL models, with a dataset of 871 test data and 218 data for testing from the WAG "Forum DTC Riau". These test results can be further improved through the proposed hybrid model.

The testing with the hybrid LSTM BiLSTM algorithm referred to previous studies [9, 13, 17, 18]. This testing still utilized the CNN BiLSTM model but with changes in the Preprocessing and Feature Extraction stages, as well as the hybrid architecture with different Activation, Dropout, Filter, and kernel size of 3, as shown in Table II and Fig. 1.

The model was trained for 20 epochs, with a batch size of 256 for testing. The training accuracy and loss, as well as the validation accuracy and loss, were calculated and represented visually, as shown in Fig. 3. The performance evolution of the proposed model during the training and testing phases was reflected in the visual representation.

The training results of the hybrid CNN-BiLSTM model on both datasets show the performance evolution over 20 training epochs. For the DTC dataset, the accuracy was initially approximately 0.7382 and increased to 0.7681 in the first epoch, with a val_loss of approximately 1.2490. Subsequently, the accuracy remains constant at 0.7681. A significant improvement in accuracy (0.9882) occurred at epoch 7 and continued to increase to approximately 0.9977 during epochs 8–10. From epochs 11–20, the accuracy remained stable with slight fluctuations, indicating model convergence. The val_accuracy at the end of training was approximately 0.8716 (see Fig. 3 DTC).

In the PS dataset, the initial accuracy was approximately 0.8210 with a val_loss of 0.9867. In epoch 2, the accuracy increased to 0.8356 with a val_loss of 0.7486. In the subsequent epochs (epochs 3 to 5), the accuracy and val_loss continued to improve, reaching an accuracy of approximately 0.9948 with a val_loss of 0.6281. However, after epochs 6 to 20, the accuracy tended to stabilize with slight fluctuations, and val_loss slightly increased. The val_accuracy at the end of training was approximately 0.8813 (see Fig. 3 PS).

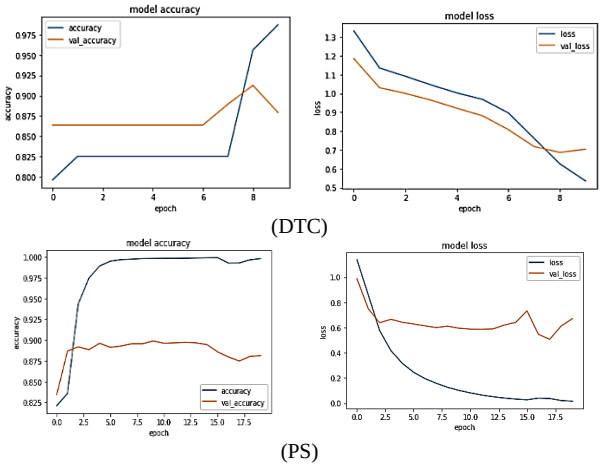

Fig. 3. Accuracy vs. epochs, loss vs epochs plot obtained from the proposed model.

The model's performance throughout both the training and testing stages is clearly reflected in the visual representation provided. After training, the proposed model achieved an impressive 98% accuracy rate for the test data. The assessment of the trained model on the

testing data produced a confusion matrix, as depicted in Fig. 4. The hybrid CNN BiLSTM model attained the highest accuracy on the PS dataset (0.8813) and slightly lower accuracy on the DTC dataset (0.8716), possibly indicating a learning plateau. The difference in convergence speed between the two datasets stemmed from their distinct characteristics.

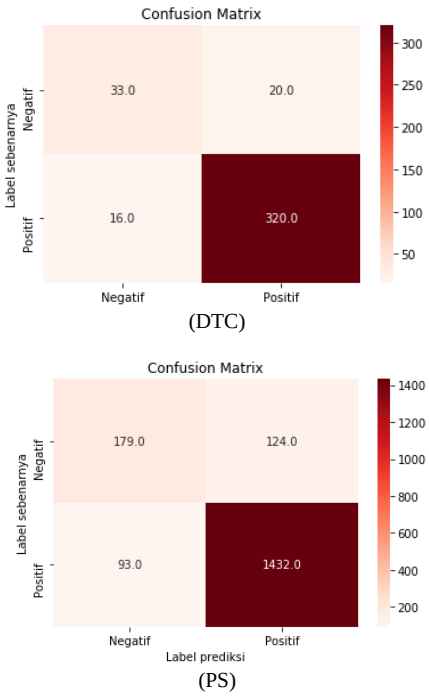

Fig. 4. Confusion matric for the proposed model.

Based on the aforementioned test results, for the single DL model, the CNN and BiLSTM exhibited better accuracy on the tested datasets, as shown in Table IV. The BiLSTM model achieved the highest accuracy, approximately 0.81% on the DTC dataset and 0.82 on the PS dataset. Precision, recall, and F1-Scores on the PS dataset were lower than on the DTC dataset, with precision around 0.53% (PS) and 0.85% (DTC), recall approximately 0.49% (PS) and 0.79% (DTC), and F1-Score about 0.51% (PS) and 0.80% (DTC). The LSTM model also achieved an accuracy of approximately 0.81% on both datasets, with similar precision, recall, and F1-Scores.

The Simple RNN model had an accuracy of approximately 0.75% on the DTC dataset and 0.80 on the PS dataset. However, this model exhibited a higher recall value on the PS dataset (0.59%) than on the DTC dataset (0.73%). The CNN model had the lowest performance on both datasets, with an accuracy around 0.70 on the DTC dataset and 0.75 on the PS dataset. The precision, recall, and F1-Scores were also generally lower than those of the other models on both datasets. These results indicate better performance compared to the study conducted by [9], which achieved an accuracy of 77.4%.

The hybrid CNN BiLSTM model on both the DTC and PS datasets demonstrated high accuracy, with a value of 0.88, confirming its capability to classify text sentiment. However, a precision value of 0.76 indicates potential false positive results. The relatively high recall values (0.82% for DTC and 0.79 for PS demonstrate the model's ability to identify most of the positive sentiments that should be detected. The F1-Scores of 0.79% the DTC and 0.78% for PS, it indicates a good balance between precision and recall, making this model a strong choice for text sentiment classification.

Testing of the hybrid CNN BiLSTM model produced varying accuracy results. This demonstrates that the model consistently outperformed the other acquired findings. The accuracy of the proposed model (88%) surpassed that of the single DL model tested in Table IV. The performance of the hybrid model in this testing was superior to existing models [9, 13, 17, 18] in terms of accuracy, as presented in Table I, when compared to the proposed model.

The findings of this study demonstrate that the combination of optimization strategies and regularization techniques, particularly the utilization of dropout, makes a noteworthy contribution to enhancing the accuracy of the tested model. The adoption of the Adam optimizer with a learning rate (lr) of 0.0001 and the implementation of regularization techniques, including dropout and L2 regularization, are crucial elements in tackling overfitting concerns and fortifying model stability. The integration of a dropout layer plays a pivotal role in diminishing the model's reliance on features.

Overall, the sentiment analysis model exhibited significant performance between the training and testing processes, leading to overfitting. Several main causes can be speculated to explain overfitting on the testing dataset. Firstly, some text messages couldn't be translated and were truncated, reducing the overall information. This can affect the lexicon-based labelling process, leading to incorrect sentiment labelling due to the diminished information in the truncated text messages.

Consequently, there are biases in the labelling process where some highly positive sentiment words also exist in the negative sentiment, making the sentiment analysis model "confusing" in generalizing unseen data. Secondly, biased labelled data can also influence the training data, where some mislabeled text messages do not fully represent the features of unseen data. Another consideration is the configuration of the sentiment analysis model with parameters that have not been well-optimized.

V. Conclusion

This study presents a unique hybrid CNN-BiLSTM model tailored exclusively for analyzing WAG data. Leveraging the strengths of both CNN and BiLSTM, this model efficiently represents both the short-term and long-term interdependence found in sequential data. The architecture includes preprocessing, feature extraction, and a hybrid structure with activation, dropout, filter, and a kernel size of 3 with multiple layers. The experimental results show that the suggested CNN-BiLSTM model achieves an astounding accuracy rate of 88%. Notably, as compared to the standalone single Bidirectional LSTM model, the CNN-BiLSTM model improves by 7 points, demonstrating its usefulness in handling WAG data.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Susandri primarily conducted this research with development of the proposed model; Sarjon Defit and Muhammad Tajuddin helped repair the paper; all authors ultimately agreed on the final version.

REFERENCES

[1] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, no. Ml, pp. 42689–42707, 2020. doi: 10.1109/ACCESS.2020.2976744

[2] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, pp. 6629–6643, 2020. doi: 10.11591/IJECE.V10I6.PP6629-6643

[3] W. Fang, H. Luo, S. Xu, P. E. D. Love, Z. Lu, and C. Ye, "Automated text classification of near-misses from safety reports: An improved deep learning approach," *Adv. Eng. Informatics*, vol. 44, no. March 2019, 101060, 2020. doi: 10.1016/j.aei.2020.101060

[4] Z. Liu, C. Lu, H. Huang, S. Lyu, and Z. Tao, "Hierarchical Multi-granularity attention- based hybrid neural network for text classification," *IEEE Access*, vol. 8, pp. 149362–149371, 2020. doi: 10.1109/ACCESS.2020.3016727

[5] H. Yang, L. Luo, L. P. Chueng, D. Ling, and F. Chin, "Deep learning and its applications to natural language processing," in *Deep Learning: Fundamentals, Theory and Applications*, 2019, pp. 89–109.

[6] R. Joshi, P. Goel, and R. Joshi, "Deep learning for hindi text classification: A comparison," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 2020, pp. 94–101. doi: 10.1007/978-3-030-44689-5_9

[7] Q. Li *et al.*, "A survey on text classification: From shallow to deep learning," *IEEE Trans. NEURAL NETWORKS Learn. Syst.*, vol. 31, no. 11, pp. 1–21, 2020.

[8] F. Zaman, M. Shardlow, S. Hassan, and N. Radi, "HTSS : A novel hybrid text summarisation and simplification architecture," *Inf. Process. Manag.*, vol. 57, no. 6, 102351, 2020. doi: 10.1016/j.ipm.2020.102351

[9] K. Pasupa, T. Seneewong, and N. Ayutthaya, "Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding , POS-tag , and sentic features," *Sustain. Cities Soc.*, vol. 50, no. 7, 101615, 2019. doi: 10.1016/j.scs.2019.101615

[10] K. Miok, D. Nguyen-Doan, B. Škrlj, D. Zaharie, and M. Robnik-Šikonja, "Prediction uncertainty estimation for hate speech classification," *Statistical Language and Speech Processing*, pp. 286–298, 2019.

[11] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate speech detection using word embedding and deep learning in the arabic language context," in *Proc. 9th International Conference on Pattern Recognition Applications andMethods (ICPRAM 2020)*, 2022, pp. 453–460. doi: 10.5220/0008954004530460

[12] A. Garain, "The titans at semeval-2019 task 6: Offensive language identification, categorization and target identification," in *Proc. 13th International Workshop on Semantic Evaluation (SemEval-2019)*, 2019, pp. 759–762.

[13] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, "Applied sciences Bi-LSTM Model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, 5841, 2020.

[14] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, "Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification," *IEEE Access*, vol. 8, pp. 77060–77072, 2020. doi: 10.1109/ACCESS.2020.2989428

[15] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," *Comput. Sci. Rev.*, vol. 38, 100311, 2020. doi: 10.1016/j.cosrev.2020.100311

[16] S. Kumar, C. Akhilesh, K. Vijay, and B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," *Vis. Comput.*, vol. 38, no. 12, pp. 4095–4109, 2021. doi: 10.1007/s00371-021-02283-3

[17] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020. doi: 10.1109/ACCESS.2020.2982538

[18] U. Naqvi, A. Majid, and S. A. L. I. Abbas, "UTSA : Urdu text sentiment analysis using deep learning methods," *IEEE Access*, vol. 9, pp. 114085–114094, 2021. doi: 10.1109/ACCESS.2021.3104308

[19] J. Gaglani, Y. Gandhi, S. Gogate, and A. Halbe, "Unsupervised whatsapp fake news detection using semantic search," in *Proc. International Conference on Intelligent Computing and Control Systems, ICICCS 2020*, 2020, pp. 285–289. doi: 10.1109/ICICCS48265.2020.9120902

[20] H. T. Assaggaf, "A discursive and pragmatic analysis of whatsapp text-based status notifications," *Arab World English J.*, vol. 10, no. 4, pp. 101–111, 2019. doi: 10.24093/awej/vol10no4.8

[21] Y. Zhou, Q. Zhang, D. Wang, and X. Gu, "Text sentiment analysis based on a new hybrid network model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, 2022.

[22] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks," *J. Big Data*, vol. 6, no. 1, 2019. doi: 10.1186/s40537-019-0246-8

[23] H. Aljuaid, R. Iftikhar, S. Ahmad, M. Asif, and M. Tanvir Afzal, "Important citation identification using sentiment analysis of in-text citations," *Telemat. Informatics*, vol. 56, 101492, 2021. doi: 10.1016/j.tele.2020.101492.

[24] N. Chintalapudi, G. Battineni, M. Di Canio, G. G. Sagaro, and F. Amenta, "Text mining with sentiment analysis on seafarers' medical documents," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, 100005, 2021. doi: 10.1016/j.jjimei.2020.100005

[25] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis," *Multimed. Tools Appl.*, vol. 78, no. 18, pp. 26597–26613, 2019. doi: 10.1007/s11042-019-07788-7

[26] S. Soumya and K. V Pramod, "Hybrid deep learning approach for sentiment classification of malayalam tweets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 891–899, 2022.