CHAPTER 1

INTRODUCTION

1.1 Background

Hate speech refers to public expressions or statements that promote hatred and offensive rhetoric directed at specific individuals or groups. Common targets of hate speech include topics such as race, religion, gender, and sexual orientation [1]. Today, particularly on social media, hate speech almost always involves offensive language. The ongoing spread of such hate can result in discriminatory behavior, fostering stigma among readers who may believe there are no repercussions for their actions. Hate speech containing offensive words and phrases often exacerbates social conflicts, stirring emotions and triggering reactions in people [2]. During online interactions, people often feel the need to defend themselves and react aggressively, leading to expressions of hatred. One widely used platform in this context is Twitter (X), which, like other social media platforms, allows users to post content known as tweets and share them with followers via retweets. Despite its popularity, Twitter (X) has become a breeding ground for the spread of hate and misinformation, largely due to its viral reach and the anonymity it offers users [3, 4].

The study of hate speech detection [5] in the context of legislative elections, such as the 2024 elections for members of the House of Representatives of the Republic of Indonesia election on Twitter, predominantly reflects negative sentiment. The study employed CNN and LSTM models across three data ratios: 90:10, 80:20, and 70:30. It was found that these data ratios had a significant impact on the models' performance. Additionally, implementing an oversampling process improved the models' effectiveness. The CNN model showed better overall performance compared to the LSTM model. At the 80:20 ratio, the CNN model using Word2Vec extraction features achieved the highest accuracy, F1-score, precision, and recall. The CNN model achieved an accuracy of 93.27%, an F1-score of 93.19%, a precision of 93.52%, and a recall of 92.73%. Overall, the CNN model surpasses the LSTM model in both accuracy and performance. Future studies should investigate alternative feature extraction techniques, incorporate multiple models, and assess performance on larger and more diverse datasets to determine optimal combinations and further improve model effectiveness.

Subsequent studies have explored the integration of different models to improve hate speech classification, particularly in the context of Indonesian elections [6]. Hate speech classification with an election topic in Indonesia was conducted using the IndoBERT model combined with an RNN layer called BiLSTM. The dataset used was obtained from data crawling using the X API. The first dataset, obtained from Alfina et al., is a binary-labeled dataset with categories for Hate Speech (HS) and non-Hate Speech (non-HS), containing 713 data points. The second dataset is a multi-label dataset with twelve categories: HS, HS_Individual, HS_Group, HS_Religion, HS_Race, HS_Physical, HS_Gender, HS_Other, HS_Weak, HS_Moderate, HS_Strong, and Abusive, totaling 13,169 data points. For this study, the multi-label dataset was consolidated into two categories: HS and non-HS. The data was classified accordingly, and these labels were used for model training. The evaluation results of this model achieved an accuracy of 88.5% for BERT base, 88.6% for BERT+BiLSTM, and 88% for CNN. These results are quite good for binary hate speech classification and suggest that further research into multi-label classification would be interesting.

Based on previous study, the results from binary hate speech classification models show very good performance, with high accuracy on CNN based 93.27% [5], 88.5% for BERT base, 88.6% for BERT+BiLSTM, and 88% for CNN [6], indicating that binary models are effective for positive and negative classification. The other study [7] on hate speech detection using tweets, deep learning models such as CNN, GRU, and their combinations (CNN+GRU, GRU+CNN) were evaluated. The study [7] identified the optimal data split as 90:10 and highlighted the effectiveness of Unigram+Bigram n-grams with a feature size of 5000. With a dataset of 63,984, with a data sharing ratio of 90:10 and with varying sentence lengths, resulting in the CNN model, particularly when combined with feature expansion techniques like the IndoNews Top 10 corpus, achieved the highest accuracy of 88.79%. However, this study is limited by the use of an Indonesian language dataset, and the average sentence length of the dataset used is unknown, so future research should include additional feature extraction methods, such as TF-IDF, and explore datasets in other languages for broader validation.

In another study [8], multilabel hatespeech classification was conducted using the indoBERT-Lite Base and BiLSTM-CNN methods optimized with Grid Search Hyperparameter. Data was collected from Github and processed through several steps, such as removing irrelevant elements, replacing non-standard words, tokenizing, and stemming. The model utilized a batch size of 90, a dropout rate of 0.2, and 30 neurons. The results showed an accuracy of 72.61%, an improvement compared to the Random Forest Decision Tree (RFDT) and Label Power-set (LP) methods, which achieved an accuracy of 66.12%. This study suggests further development by increasing parameters such as dropout rate and neuron units, as well as enhancing the number of epochs to improve accuracy and data quality.

Another study [9], the RFDT, BiLSTM, and BiLSTM with a pre-trained BERT model were used for multilabel classification of hate speech on Twitter in Indonesian, English, and mixed languages up to the type, category, and level. The process involved data transformation using Classifier Chains, Label Powerset, and Binary Relevance, as well as feature extraction with TF-IDF. Experiments were conducted under various preprocessing scenarios, including without translation, without stemming, and without stopword removal.

The best result was achieved using the RFDT method with Classifier Chains, without translation, stemming, or stopword removal, resulting in an accuracy of 76.12%. The study revealed that translation, stemming, and stopword removal were less effective, while label dependencies significantly influenced the results. It is recommended to enrich the dataset to enhance features and apply more complex deep neural network methods with hyperparameter tuning.

Multiclass classification is interesting to investigate for several key reasons. First, multiclass classification allows for more in-depth and detailed analysis compared to binary classification. By classifying text into several subtypes of hate speech, such as Ethnicity, Religion, Race, and Inter-group, researchers can gain a more nuanced understanding of the variations and contexts of hate speech. Second, multiclass classification can handle a wider range of data variations, aiding in the understanding of different forms of hate speech that might not be detected with binary models. Third, integrating BERT and deep learning models with multiclass classification can leverage better contextual understanding and feature extraction, enhancing the accuracy and strength of the model. Last, multiclass classification offers a richer and more nuanced approach to understanding and classifying hate speech, making it a highly compelling area of research [10].

On another study [11], research on hate speech detection was also influenced by the text length used. The study focused on abusive and hate speech in Indonesian local languages, such as Javanese, Sundanese, and Minangkabau. This research utilized various models, including SVM, Multinomial Naive Bayes, and Random Forest Decision Trees (RFDT), along with transformation methods like Binary Relevance, Classifier Chain, and Label Powerset. The dataset consisted of ”long text” (over 100 characters) with 2,066 rows and ”short text” (100 characters or less) with 4,472 rows. The results indicated that SVM combined with the Classifier Chain method and unigram features provided the best performance for Javanese and Sundanese datasets. Meanwhile, RFDT with similar methods achieved the highest F1-score of 80.75% for the Minangkabau dataset. Despite these advances, the study encountered challenges in accurately classifying invective categories and addressing imbalanced datasets, particularly for languages like Madurese and Minangkabau. To improve model performance, the study recommended strategies such as data balancing and hyperparameter tuning.

Integrating BERT and other deep learning models with multiclass classification can significantly enhance contextual understanding and feature extraction, improving the accuracy and robustness of these models. Future research should prioritize the integration of multiple models, advanced feature extraction methods, and diverse datasets to further refine these approaches. Table 1.1 shows highlights the strengths and weaknesses of previous research, and provides recommendations for future research. This positions multiclass classification as a promising and comprehensive framework for addressing the complexities of hate speech.

Table 1.1 provides a comparative analysis of different text classification methods, highlighting their strengths and weaknesses.

Table 1.1: Hate Speech Advantages and Disadvantages

Reference	**Focus**	**Method**	**Advantages**	**Disadvantages**
J. Forry Kusuma and A. Chowanda [6]	Hate Speech Detection	IndoBERT and BiLSTM	Integration of IndoBERT with BiLSTM achieved high accuracy.	Limited to binary classification, losing valuable information from multi-label datasets. Using data with varying sentence lengths, whereas the data contains long text (over 100 characters) comprising 2,066 rows and ”short text” (100 characters or less) comprising 4,472 rows, which can be explored further.
K. U. Wijaya and E. B. Setiawan [7]	Hate Speech Detection using a Combined Model	Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and hybrid models combining CNN-GRU and GRU-CNN for classification.	Achieved high accuracy (88.79%) with CNN and Unigram+Bigram n-grams.	Did not explore other feature extraction methods like TF-IDF and did not mention sentence length, also did not experiment with sentence length.

Reference	**Focus**	**Method**	**Advantages**	**Disadvantages**
A. D. Asti, I. Budi, and M. O. Ibrohim [11]	Multi-label Classification for Hate Speech and Abusive Language in Indonesian Local Languages	Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), and Random Forest Decision Tree (RFDT) utilizing transformation techniques like Binary Relevance (BR), Classifier Chains (CC), and Label Powerset (LP).	The study's focus on multi-label classification allows for more detailed categorization of hate speech and abusive language. By concentrating on Indonesian local languages, the study addresses the nuances and specificities of hate speech in different regional contexts.	The study faced challenges in accurately classifying invective categories and balancing the datasets, particularly for languages like Madurese and Minangkabau. The study recommends data balancing and hyperparameter tuning to enhance model performance. Also ignored the element of sentence length, but in the example sentence, the sentence length is 5 words. It is possible that there are shorter and longer sentences.
Hulliyah, Khodijah and Muzayyanah, Fenty Eka and Setyawan, Bayu Aji[8]	Multilabel classification of hate speech severity	indoBERT-Lite Base and BiLSTM-CNN	Achieved 72.61% accuracy, improved from 66.12% with RFDT and LP methods	Requires further optimization of dropout rate, neuron units, and epochs to enhance accuracy

Reference	Focus	Method	Advantages	Disadvantages
Hendrawan, Rahmat and Adiwijaya and Al Faraby, Said [9]	Multilabel hate speech classification on Twitter in Indonesian, English, and mixed languages	RFDT, BiLSTM, and BiLSTM with pre-trained BERT	Achieved 76.12% accuracy without translation, stemming, or stopword removal	Translation, stemming, and stopword removal were less effective; dataset enrichment and advanced methods needed for improvement

On Table 1.2 highlights a comprehensive comparison of the advantages and disadvantages of various text classification methods, providing valuable insights into their strengths and limitations for different applications.

Table 1.2: Advantages and Disadvantages of Models for Hate Speech Detection

Model	Advantages	Disadvantages
BERT	Contextual understanding in both directions with high accuracy.	Requires high computational power and slow inference.
CNN	Fast and effective in detecting local patterns (n-grams).	Does not deeply understand word sequences.
LSTM	Captures word sequences with long-term memory.	Slow due to sequential nature and prone to overfitting.
BiLSTM	Captures bidirectional context for better accuracy.	Slower and requires more memory.
RNN	Good for simple sequential data with short dependencies.	Suffers from vanishing gradient issues, unsuitable for complex data.
GRU	Faster and lighter than LSTM.	Less effective for complex contexts compared to BiLSTM.
FastText	Fast and efficient for small datasets.	Does not understand sequence or deep contextual meaning.
SVM	Simple and effective for small datasets.	Ineffective for large or high-dimensional text data.
Random Forest	Robust against overfitting on small datasets.	Ineffective for high-dimensional text data.

Traditional models like RNN are unsuitable because they suffer from vanishing gradient issues, making it hard to capture long-term dependencies in text [12]. While GRU is faster and lighter than LSTM, its performance is suboptimal for complex multilabel classification tasks [13]. FastText is fast and efficient for small datasets but uses static word representations, which fail to capture the contextual meaning and sequence of words [14]. Machine learning-based models such as SVM and Random Forest are suitable for small, low-dimensional datasets but are inefficient for high-dimensional text data requiring complex understanding [15, 16]. Additionally, CNN without integration with a contextual model like BERT has limitations in understanding word sequences, making it less accurate for tasks requiring full context [14].

BERT (Bidirectional Encoder Representations from Transformers) demonstrates superior contextual understanding thanks to its bidirectional approach and the utilization of transformers, along with the effectiveness of transfer learning, which allows its application across various NLP tasks with outstanding performance [17]. BERT excels at capturing bidirectional context and subtle meanings in long and complex sentences, owing to its ability to maintain long-term dependencies. However, this model is complex and requires high computational resources, such as GPUs with large memory, making it slow in training and inference. BERT can effectively handle varying sentence lengths and is not bound to a specific sentence length like LSTM, whereas CNN tends to require pooling layers to address this issue [18]. CNN (Convolutional Neural Network) is better suited for capturing local features such as phrases or n-grams and performs well on short texts, but it is less optimal for long sentences due to its limitations in handling global dependencies [19]. CNN excels in inference speed and in sentiment analysis, where it effectively captures dominant information. However, it is not ideal for complex contextual understanding. LSTM (Long Short-Term Memory) is capable of capturing long-term dependencies, although its processing is unidirectional, making it less efficient than BERT in bidirectional contexts. LSTM performs better than CNN on long texts and excels at understanding long-term semantic dependencies, making it suitable for lengthy text data [20]. BiLSTM (Bidirectional LSTM), as an enhancement of LSTM, processes text in both directions and is better at understanding complex contexts, although it still slightly lags behind BERT in performance and speed [21].

Hybrid models combining BERT (Bidirectional Encoder Representations from Transformers) with other deep learning approaches have been shown to yield good results in the context of Natural Language Processing [8, 9]. BERT, as a transformer-based model pre-trained on two main tasks—masked language modeling and next sentence prediction—enables richer and more contextual word representations. Therefore, using BERT early in a hybrid model plays a crucial role in generating better text representations and reducing errors in context understanding that may occur in tasks such as sentiment analysis, information extraction, and entity recognition[17]. After the initial processing stage using BERT, other deep learning layers can be used to process and combine these contextual features for more complex specific tasks, such as class prediction or deeper pattern recognition. This approach allows the model to leverage BERT's strengths in understanding the semantic context of the text, while other deep learning models optimize the results based on broader data [6, 8, 9].

Buddy Media's research [22] aligns with similar research by Track Social [23] in a study of 100 popular brands on Twitter. Track Social also found that the ideal Tweet length is around 100 characters. Their analysis saw a spike in retweets among tweets ranging from 71 to 100 characters—what they call "medium" length tweets. These medium-length tweets have enough characters for the original tweeter to say something valuable and for the retweeter to add commentary as well. Based on Haryadi research [24], text usage on twitter is mostly in the length of 140 characters. Text in 40 characters shows a tendency to be able to spread information compared to text with a shorter length. In recent years, Twitter increased the character limit for Tweets from 140 to 280 characters. X now offers 280 characters for your tweets. However, Twitter Blue users can send tweets up to 10,000 characters long. In recent years, Twitter increased the character limit for Tweets from 140 to 280 characters. X now offers 280 characters for your tweets. However, Twitter Blue users can send tweets up to 4,000 characters long [25]. Currently, there are no journals or studies that specifically discuss the division of text into short text and long text categories, especially in the context of data analysis or social media. This topic is still a potential research gap to be explored further, given the importance of understanding the influence of text length on user engagement and information dissemination.

So far, research on the effect of sentence length is still relatively minimal compared to cross-domain and cross-language text classification [26]. This may be due to the common assumption that differences in text length distribution do not significantly affect classification performance as long as the text content is similar. However, the study conducted by [27] demonstrates that this assumption does not fully hold true. The research highlights a notable drop in accuracy across several text classification techniques, including BoW, CNN, and BERT, when these models are applied to predict texts with lengths that differ from the training data. Amplayo et al. observed that the classification accuracy could decline significantly, with reductions ranging from 3% to 17.5%, depending on the dataset and the model employed. These findings underline the importance of considering text length as a critical factor in designing and optimizing deep learning-based text classification models. Research by Faraby [26], specifically investigated the impact of cross-length conditions by creating a specialized dataset and evaluating it with various widely used classification models. The results indicated that differences in text length distribution between training and testing data could significantly influence model performance. When transferring from long to short texts, the average F1-score dropped by 14% across all models, while transferring from short to long texts led to an average decrease of 9%. Other Research by Pambudi [28], for binary datasets, it can be concluded that sentence length affects the performance of

SVM and CNN algorithms when Word2Vec is used for feature weighting. However, when TF-IDF weighting is combined with the SVM algorithm, sentence length has no significant impact on performance.

Based on previous research [26–28], sentence length is a crucial factor in text classification models, as differences in length distribution between training and testing data have been shown to significantly affect model performance. These findings emphasize the importance of considering text length in model design and optimization to improve the accuracy and effectiveness of classification results.

This study seeks to explore the differences between independently trained BERT models and those trained in conjunction with other deep learning models for detecting hate speech in the Indonesian language. Ultimately, the research aims to improve the understanding of hate speech detection techniques on Indonesian social platforms and contribute to the development of BERT-specific models based on the findings of this study.

Combining BERT with other deep learning models (CNN, LSTM, BiLSTM) provides a hybrid approach that leverages the strengths of each model. BERT offers rich contextual features as embeddings, CNN improves classification efficiency, and LSTM/BiLSTM ensures that sequence and deep context are preserved. This combination achieves an optimal balance between accuracy and efficiency for handling hate speech multilabel classification tasks on large and complex datasets, such as Twitter during elections [13, 14].

1.2 Theoretical Framework

BERT (Bidirectional Encoder Representations from Transformers) demonstrates superior contextual understanding thanks to its bidirectional approach and the utilization of transformers, along with the effectiveness of transfer learning, which allows its application across various NLP tasks with outstanding performance [17] BERT excels at capturing bidirectional context and subtle meanings in long and complex sentences, owing to its ability to maintain long-term dependencies. However, this model is complex and requires high computational resources, such as GPUs with large memory, making it slow in training and inference. BERT can effectively handle varying sentence lengths and is not bound to a specific sentence length like LSTM, whereas CNN tends to require pooling layers to address this issue [18]. CNN (Convolutional Neural Network) is better suited for capturing local features such as phrases or n-grams and performs well on short texts, but it is less optimal for long sentences due to its limitations in handling global dependencies [19]. CNN excels in inference speed and in sentiment analysis, where it effectively captures dominant information. However, it is not ideal for complex contextual understanding. LSTM (Long Short-Term Memory) is capable of capturing long-term dependencies, although its processing is unidirectional, making it less efficient than BERT in bidirectional contexts. LSTM performs better than CNN on long texts and excels at understanding long-term semantic dependencies, making it suitable for lengthy text data [20]. BiLSTM (Bidirectional LSTM), as an enhancement of LSTM, processes text in both directions and is better at understanding complex contexts, although it still slightly lags behind BERT in performance and speed [21].

Given the strengths and weaknesses of each model, combining BERT with other deep learning architectures such as CNN, LSTM, or BiLSTM can leverage their complementary advantages. For instance, CNN's efficiency in capturing local features can enhance BERT's global contextual understanding, particularly in short texts, while LSTM and BiLSTM can strengthen BERT's ability to capture sequential and long-term dependencies in lengthy texts. This integration allows the model to address varying text characteristics more effectively, ensuring robust performance across diverse datasets. By utilizing the strengths of each architecture, the combined model can overcome individual limitations, providing a more versatile and accurate approach to complex tasks like hate speech detection.

1.3 Conceptual Framework/Paradigm

In recent years, BERT (Bidirectional Encoder Representations from Transformers) has emerged as one of the leading models for text classification, including hate speech detection. The primary strength of BERT lies in its ability to understand bidirectional context in text through its transformer architecture. A study demonstrated that integrating BERT with a BiLSTM layer achieved an accuracy of 88.6% for hate speech detection on Twitter during the 2024 Indonesian General Election [6]. Additionally, BERT's transfer learning capability enables its application across various NLP tasks, making it a flexible and powerful tool. However, using BERT requires high computational resources and longer training times compared to other models [17].

Other models like CNN (Convolutional Neural Network) have also shown good performance in text classification tasks. In one study, a CNN model combined with Word2Vec features achieved the highest accuracy of 93.27% for sentiment analysis related to Indonesia's Legislative Elections[5]. CNN's fast inference speed and ability to capture local features make it ideal for short texts, although it is less optimal for long texts due to its limitations in understanding global dependencies

Further research could explore the integration of BERT with other models, such as CNN or BiLSTM, to improve detection accuracy.[10, 19].

1.4 Statement of the Problem

In previous studies, the primary focus has been on identifying hate speech using binary classification methods, which typically categorize content as either hate speech or not [29–31]. These methods have utilized various machine learning and deep learning models, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). However, these binary approaches may not be sufficient to capture the complexity and nuances of hate speech, which can vary significantly in form and target [29]. Additionally, most research [32–34] has concentrated on developing methods for hate speech detection, often focusing on single-label classification techniques. These techniques classify text into one category, either hate speech or not, based on predefined labels. However, the effectiveness of these methods can be limited when dealing with more complex and multilayered hate speech content, which may require a more nuanced approach [30]. In previous studies [7, 11], sentence length has not been considered as a testing variable affecting the effectiveness of hate speech detection. Most research has focused primarily on model selection and classification methods without considering that variations in sentence length may impact a model's accuracy in identifying hateful content. Shorter sentences may lack sufficient context, making it more challenging for models to accurately detect hate, while longer sentences can introduce additional complexity that may confuse classification models. Therefore, in this study, an analysis based on sentence length will be used to measure the effectiveness of the model in handling variations in text length, aiming to achieve more accurate and contextually relevant detection results. Given these limitations, it is crucial to explore more sophisticated models that can handle multilabel classification. This approach can classify text into multiple categories simultaneously, providing a more detailed understanding of the content and its context [31]. Furthermore, combining different models, such as BERT with other deep learning architectures, can leverage their respective strengths and improve the overall performance of hate speech detection systems, as well as understand the effectiveness of detection based on sentence length.

1.5 Objective and Hypotheses

This research aims to categorize hate speech based on specific contexts such as race, gender, and religion, as well as identify effective combinations of ensemble methods for detecting hate speech on the Twitter (X) platform using text-based data. Additionally, it is expected to enhance hate speech detection by adopting an ensemble approach that combines the Bidirectional Encoder Representations from Transformers (BERT) method with other deep learning models. A key objective of this research is to analyze performance based on experiments involving sentence length, as longer sentences may present unique challenges and nuances in hate speech detection.

The expected outcome of this research is to provide deeper insights into the patterns and trends of hate speech on social media and contribute to the development of more effective and context-sensitive detection algorithms. The combination of BERT and Bi-LSTM is anticipated to yield better results for understanding the characteristics and behavioral patterns of users prone to spreading hate speech. This combination leverages the strengths of both models: BERT's ability to comprehend global context and Bi-LSTM's capability to capture bidirectional dependencies within sequences. By integrating these two models, the ensemble can more accurately identify and analyze subtle patterns of hate speech text. Additionally, the incorporation of Bi-LSTM enhances context understanding by retaining sequential information effectively. Given that hate speech often targets sensitive topics such as religion and intergroup issues on social media, the ensemble model is anticipated to effectively classify and identify the most common and relevant types of hate speech within the context of the Twitter (X) platform. Thus, the combination of BERT and Bi-LSTM is expected to provide superior performance due to their complementary strengths in capturing both global and sequential contextual information. Furthermore, by analyzing the impact of sentence length on detection accuracy, this research aims to determine how variations in text length influence model performance. This analysis is expected to reveal whether shorter texts, which may lack sufficient context, are more challenging for the model to classify accurately compared to longer, more detailed texts. Through these insights, this study aims to refine hate speech detection algorithms to handle the diversity in sentence lengths, ultimately providing a more nuanced understanding of hate speech patterns across varied text lengths. Thus, the combination of BERT and Bi-LSTM is expected to provide superior performance due to their complementary strengths in capturing both global and sequential contextual information.

1.6 Assumption

The study assumes that BERT can be optimized using a combination with deep learning, and sentence length will have a significant impact on text classification performance.

1.7 Scope and Delimitation

This research focuses on multilabel classification using BERT combined with deep learning for social media data, specifically tweets related to the Indonesian Election 2024 with testing on sentence length. The study will not cover other social media platforms or topics outside the election context.