IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 4, August, 2023, pp. 929-936
Journal homepage: http://iieta.org/journals/ria

Enhancing Cyberbullying Detection on Indonesian Twitter: Leveraging FastText for Feature Expansion and Hybrid Approach Applying CNN and BiLSTM

Muhammad Alfi Syahri Nasution, Erwin Budi Setiawan*

Informatics, School of Computing, Telkom University, Bandung 40257, Indonesia

Corresponding Author Email: erwinbudisetiawan@telkomuniversity.ac.id

https://doi.org/10.18280/ria.370413

Received: 17 June 2023
Revised: 23 July 2023
Accepted: 2 August 2023
Available online: 31 August 2023

Keywords:
cyberbullying, feature expansion, Fasttext, Hybrid Deep Learning, Convolutional Neural Network, Bidirectional Long Short-Term Memory

ABSTRACT

Cyberbullying, characterized by the transmission of threatening, intimidating, and derogatory messages via digital platforms such as Twitter, is a pervasive issue. Given the volume of approximately 867 million daily tweets, the potential scale of cyberbullying incidents is immense, underscoring the necessity for automated detection systems for such messages. However, the context-sensitive nature of tweets can pose challenges to understanding message content, particularly in languages like Indonesian with potential for significant vocabulary discrepancies. This study aims to enhance cyberbullying detection by employing feature expansion using FastText, thereby addressing vocabulary-related comprehension issues in Indonesian-language tweets. Furthermore, text classification is performed using a Hybrid Deep Learning approach, integrating Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). This hybrid model leverages the strengths of both techniques, capturing local patterns and long-range dependencies within the data. The objective of this research is to evaluate the performance yielded by the application of FastText-enhanced feature expansion and Hybrid Deep Learning to an Indonesian Twitter dataset. This focus is motivated by the high accuracy of Hybrid Deep Learning for Twitter datasets in other languages, and the limited application of such methods to Indonesian-language datasets, which predominantly use supervised learning or deep learning. Analysis of 29,085 datasets demonstrated that the combined implementation of Hybrid Deep Learning and FastText-enhanced feature expansion achieved the highest accuracy, with CNN-BiLSTM and BiLSTM-CNN scoring 80.55% and 80.35% respectively. These findings validate the significant accuracy boost provided by FastText when integrated with Hybrid Deep Learning. It is anticipated that the outcomes of this study will facilitate the accurate identification and removal of cyberbullying tweets, thereby contributing to a safer digital communication environment on Twitter.

1. INTRODUCTION

Social media has increasingly become an integral part of everyday life, seamlessly woven into the fabric of personal and professional interactions [1]. Among various platforms, Twitter has emerged as a popular choice, offering users the capability to post and read others' messages, commonly referred to as tweets. These brief updates, limited to 280 characters, serve as a snapshot into the thoughts, interests, and activities of the user. As one of the largest social media platforms, Twitter sees an immense volume of activity, with data indicating approximately 867 million tweets dispatched daily [2]. However, the content disseminated through this platform is not universally positive. Among the myriad of tweets, a significant portion harbors negative content, with cyberbullying serving as a prime example. In its broadest sense, cyberbullying encompasses the use of digital communication tools, such as social media, to send threatening, intimidating, or demeaning messages. The pervasive nature of platforms like Twitter brings a unique set of challenges in monitoring and moderating such negative interactions, making it a crucial focus area for research and intervention strategies.

Cyberbullying involves deliberate and hostile actions conducted by individuals or groups using digital communication channels, sending harmful, threatening, intimidating, or derogatory messages and comments to targeted individuals [3]. In a polling survey conducted by the Association of Indonesian Internet Service Providers (APJII), out of a total of 5,900 respondents, it was found that around 49% of people in Indonesia claimed to have experienced bullying on social media [4]. Then, the research conducted by Nikon [5] shows that the impact of cyberbullying is severe, such as depression, emotional stress, and drug use, and can lead to suicide. So, it is necessary to develop a cyberbullying detection system, especially on Twitter automatically, because many daily messages can contain cyberbullying.

In making cyberbullying messages, we often encounter the use of abbreviated words and slang words, so it is difficult to understand, and vocabulary errors occur. So, to reduce vocabulary discrepancies, feature expansion can be done using word embedding [6]. However, based on my knowledge, a lack of research has been conducted concerning detecting cyberbullying using feature expansions. One of the uses of feature expansion can be using FastText because Kaibi and Satori [7] conducted research by comparing three word embeddings (Glove, Word2Vec, FastText), each of which was combined with machine learning algorithms such as NuSVC, Random Forest, GaussianNB, LinearSVC, SGD, and

LogisticRegression concluded that using FastText as word embedding got a higher accuracy value than the other word embedding, with an F1-Score of 81.97%. FastText can outperform Word2Vec and Glove because FastText has the ability to deal with rare words or out-of-vocabulary words [8]. Thus, FastText is the right choice to be a feature expansion, especially for words that are hard to understand, like in tweets.

Many studies related to cyberbullying detection have been carried out. Recent research has implemented Hybrid Deep Learning. Joshi et al. [9] use Hybrid Deep Learning as its text classification method with CNN, BiLSTM, and Glove Embedding as its feature extraction. This study produces an accuracy value of 92%. Aldhyani et al. [10] also applied CNN and BiLSTM as Hybrid Deep Learning and Keras Embedding as feature extraction for Cyberbullying Identification. They use a dataset from Wikipedia of 11,000 data with an accuracy value is 93%. In another study, Dewani et al. [11] used a dataset from the Roman Urdu language to detect Cyberbullying by applying Hybrid Deep Learning using RNN and BiLSTM. The accuracy score using the Roman Urdu dataset is 84% with the help of hypertuning parameters.

Furthermore, research related to cyberbullying for Indonesian-language datasets has its challenges due to the habit of Indonesian people speaking various languages such as regional languages, slang, and word abbreviations which are sometimes difficult to understand. So, FastText capabilities are needed to overcome this variety. However, research developments related to cyberbullying based on Indonesian-language datasets are still limited to Supervised Learning or Deep Learning, as done by Nurrahmi and Nurjanah [12], who used a dataset from Indonesian-language Twitter with a total of 700 tweets to detect cyberbullying with the Support Vector Machine (SVM) method. Based on this research, the F1-Score result was 67%. There is also research conducted by Andriansyah et al. [13] using a Support Vector Machine (SVM). The proposed method uses a dataset of 1053 comments on Instagram, which obtains an accuracy of 79%. Then, Muzakir et al. [14] also detect cyberbullying using a Support Vector Machine (SVM) and Bag of Words (BoW) as feature extraction. They reported that the accuracy score is 76% using a dataset from Twitter with 1065 tweets. Putri et al. [15] developed Support Vector Machine (SVM) model for cyberbullying detection based on tweets from Twitter. This study reported that the accuracy value is 76.2%. Another study presented by Laxmi et al. [16] proposed a deep learning method using Convolutional Neural Networks (CNN) and Doc2Vec as feature extraction. This study uses 725 tweets from the Indonesian-language Twitter, producing an F1-Score of 65%.

The main contribution of this research is to present a combination of a Hybrid Deep Learning model and feature expansion for detecting cyberbullying in Indonesian-language Twitter because to the best of our knowledge, no one has conducted this research, and it has the potential to increase the value of accuracy in detecting cyberbullying because, based on the related research previously described, the use of Hybrid Deep Learning for Twitter datasets in other languages has high accuracy, and research related to cyberbullying detection using Indonesian-language datasets is still limited to Supervised Learning or Deep Learning and have not used feature expansion yet. Then, choosing FastText as a feature expansion can overcome problems related to misunderstanding of vocabulary because the Indonesian-language Twitter dataset has its challenges as previously explained. Therefore, various scenarios will be attempted, such as selecting the split data ratio and feature extraction using the best n-grams. A combination of Hybrid Deep Learning models, CNN-BiLSTM and BiLSTM-CNN, will also be tested. Another scenario involves feature expansion in Hybrid Deep Learning, utilizing FastText and selecting the best-performing top-ranked from the different corpus. The last scenario attempts to hypertuning the parameter of the hybrid model.

The structure of this paper will include the following sections: Section 2 will present the research methodology for cyberbullying detection using Hybrid Deep Learning and FastText as feature expansion, and Section 3 will discuss the results and discussion. Finally, Section 4 will present the conclusion of the conducted research.

2. RESEARCH METHODOLOGY

The following Figure 1 is a proposed method for the use of feature expansion and Hybrid Deep Learning for detecting cyberbullying on Twitter.

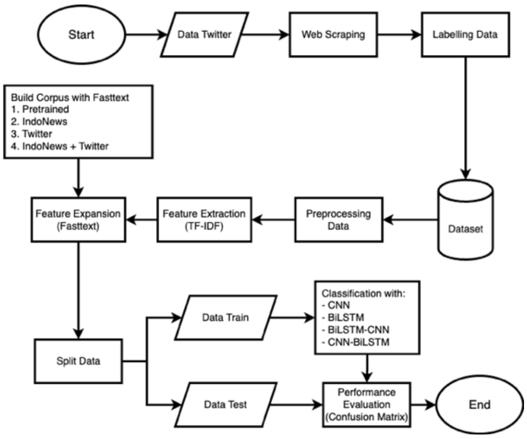

Figure 1. Proposed method

2.1 Web scraping

Web Scraping is a method used to retrieve data from data sources and produce a dataset. In this research, the data source comes from tweets on social media Twitter. The data collected from Twitter can potentially contain cyberbullying in its tweets. The API provided by Twitter will assist this Web Scraping process. In addition to this method will also perform manual data retrieval to obtain the appropriate data form. The keywords used during the data retrieval process are words that potentially contain cyberbullying, as shown in Table 1 below.

Table 1. Tweet's keyword

Keyword	Total
Tolol	7.125
Banci	3.007
Goblok	4.844
Lonte	4.078
Gendut	2.039
Bodoh	2.508
Jelek	3.756
Kontol	1.078
Bangsat	650

2.2 Data labeling

After web scraping from Twitter, the next step is data labeling. Labeling process is necessary to help the model to find out which form of data includes cyberbullying and not cyberbullying. This process uses manual labeling by three people to ensure from various perspectives and minimize bias or errors if only done by one person. Then, the final result is determined using a majority vote. Manual labeling will use Binary form. Thus, data that includes cyberbullying will be given a value of "1", and data that does not include cyberbullying will be given a value of "0". Table 2 is the result of data labeling and distribution of values.

Table 2. Distribution of labeling data

Label	Class	Total	Percentage
Cyberbullying	1	14,401	49.51%
Non-Cyberbullying	0	14,684	50.49%
Total		29,085	100%

2.3 Preprocessing data

The data contained on Twitter is text, and text is unstructured data. Therefore, data preprocessing is needed to make the resulting data more structured and smoother the classification process [17]. The following is the process of preprocessing:

1. Data cleaning removes certain elements or symbols from the data that do not affect the development of a classification model, such as numbers, username symbols (@), hashtags (#), URLs, emojis, spaces, and repeated letters.
2. Case folding is the process of standardizing sentences by converting them to lowercase. Its purpose is to facilitate the classification process by eliminating variations in letter cases within a sentence.
3. Normalization is the process of standardizing non-standard words to ensure that the model does not treat them as different words, such as "ga", "gak", "engga", and "enggak", which have the same meaning. Process normalize the words, a specialized dictionary will be used as assistance.
4. Tokenization is tokenizing or splitting a sentence into a list of words to facilitate the model's understanding of the data and enable word-by-word exploration within a sentence.
5. Stopword removal is the process of eliminating words that do not have a significant impact on the modelling process, as these words serve as connectors within a sentence, such as "and", "that", "he", and "this". This process uses dictionary stopword in the nltk library.
6. Stemming is removing prefixes and suffixes to change a word into its basic form. As a result, identical words will be identified as a single word. In this study, the stemming process uses the Sastrawi library.

2.4 Feature extraction with TF-IDF

Feature Extraction is the first stage in processing the classification of text, it is used to get a representation of a text in vector form, and each word will be given a weight. Word weighting can be done using the Term Frequency- Inverse Document Frequency (TF-IDF). TF-IDF is a combination of Term Frequency (TF) calculate the occurrence frequency of words within a document and Inverse Document Frequency (IDF) which gives weight to words. So, words often appearing in a document will have a lower value than words rarely [18]. The equation of TF-IDF is as follows:

$$W_{td} = TF_{td} \times IDF_t \quad (1)$$

$$IDF_t = \left(log \left(\frac{n}{df} \right) \right) \quad (2)$$

In the Eq. (1) and Eq. (2), W is the weight of the d document to the t word, n represents the total count of documents, and df is the number of documents containing the term of t word.

2.5 Feature expansion with FastText

Further, feature expansion will be conducted. Feature expansion is expanding the original text into a large document by adding semantics to the text to make it look more meaningful than before. Making it possible to find missing words from the tweet representation [6]. Feature expansion in this study uses FastText to determine the similarity value of these words.

FastText is Facebook's word representation library which provides 600 billion word vectors. FastText helps recognize unknown words by generating vectors and splitting them into n-gram characters [19]. FastText is a skip-gram model that converts text into vector form and represents each word as an n-gram character, making it possible to find words that are not in the corpus arising from several n-grams that make up words in the corpus [8]. This study using two types of the corpus, the pre-trained corpus provided by FastText and built corpus that using three types of data: news, tweets, and a combination of news and tweets. The pre-trained corpus is different from the built corpus that the pre-trained corpus is a word vector provided by FastText and formed into a similarity corpus to find the similarity of a word. In contrast, the built corpus is a corpus made in this study with data that comes from news, tweets, and a combination of news and tweets to make a similarity corpus that can also find a word's similarity. In this study, both types of the corpus will be tested individually to find out which corpus has a better effect on increasing the accuracy value for detecting cyberbullying. Table 3 shows the total data from each built corpus used in this study.

Table 3. FastText built corpus

Corpus	Indonews	Tweets	Tweets + Indonews
Total	142,523	29,085	171,608

FastText implementation utilizes subword embedding, which will be broken down into several letters, and each letter is represented as a bag of n-gram characters. For example, the n-gram implementation of the word "adaptation" where n has a value of 3 becomes <ada, dap, apt, pta, asi> [8]. Furthermore, after FastText gets similar words, the feature expansion process will be continued on the representation vector obtained from the previous feature extraction. Features with a vector value 0 are replaced with vector values of similar words in tweets [20]. So, FastText can predict well for words that are rarely known [8]. Table 4 shows an example of the word "Tolol", which is similar to several other words based on its similarity value.

Table 4. Example of similarity word "Tolol"

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Tolol	Bodoh	Goblok	Bego	Dungu

2.6 Classification algorithm

After the feature expansion process is carried out, the resulting vector will be input for building models with Hybrid Deep Learning using Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) to be able to detect by classifying as illustrated in Figure 1.

A Convolutional Neural Network (CNN) is part of a neural network that uses a convolution structure to extract local feature vectors [21]. CNN have a multilayer network, the output from one layer will be the input from the next layer. So, it consists of input, several hidden layers and output [22]. The complete architecture is in Figure 2.

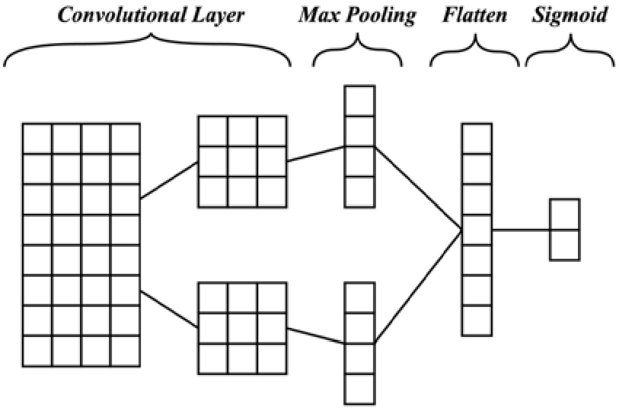

Figure 2. CNN architecture

Bi-directional Long Short-Term Memory (BiLSTM) is a directional combination of forward LSTM and backward LSTM. LSTM is a model to express a sentence by learning what to remember and forget through training. However, the LSTM model cannot encode information from back to front. Therefore, BiLSTM is needed to capture bi-directional semantic dependencies [21]. The following is the BiLSTM architecture in Figure 3.

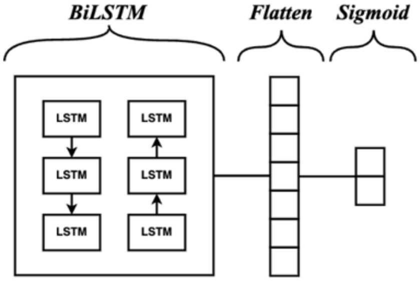

Figure 3. BiLSTM architecture

So, this study will do experiments by combining CNN and BiLSTM to build a Hybrid Deep Learning model, CNN-BiLSTM and BiLSTM-CNN, by using the strengths of each model, with CNN being able to capture local patterns and BiLSTM enable to catch past and future context and long-range dependencies [23]. A complete picture of the architecture of CNN-BiLSTM and BiLSTM-CNN is in Figure 4 and Figure 5 below.

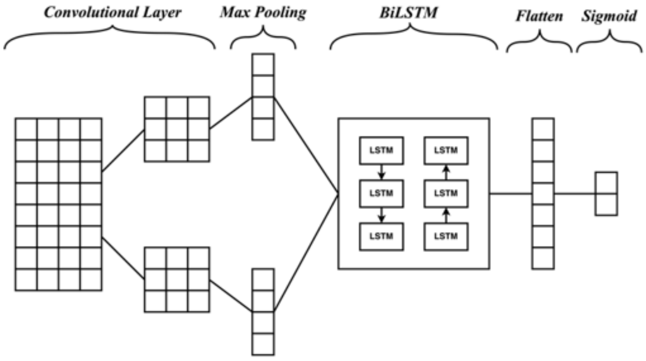

Figure 4. CNN-BiLSTM architecture

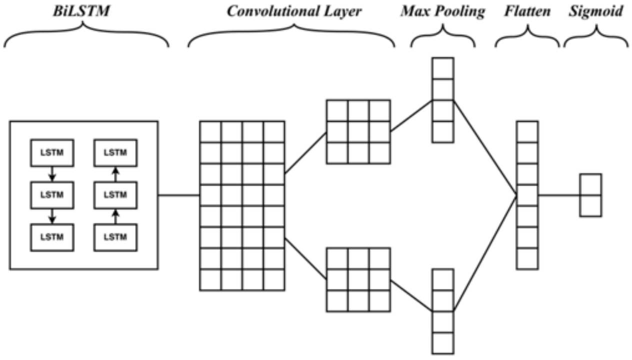

Figure 5. BiLSTM-CNN architecture

2.7 Performance evaluation

At this final step, this study evaluates the built model. Performance evaluation is carried out using the Confusion Matrix. A confusion matrix is one of the tools for measuring performance for classifications whose output can be in the form of two or more classes. There are four combinations of predicted and actual values. The four value combinations include True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [24]. From these values, it will produce one of the output values valid for performance evaluation, Accuracy. Accuracy is a measure of value to find out the ratio of the actual true value to all data using this formula in Eq. (3).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

3. RESULT AND DISCUSSION

This study conducted several experiments on four classification algorithm models, CNN, BiLSTM, CNN-BiLSTM, and BiLSTM-CNN. The results will be explained in this section and will be divided into several scenarios as follows.

3.1 Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory (BiLSTM)

In this first scenario, it will produce a model that is used as a baseline. The baseline is a reference and comparison for the following scenario to determine the effect on the accuracy value. The best baseline comes from a combination of models and Feature Extraction, which uses several TF-IDF n-grams such as Unigram, Bigram, Trigram, Unigram + Bigram or Allgram. The models to be tested in this scenario are CNN and BiLSTM with the proportion of data split 90:10, 80:20 and 70:30. In this study, the parameters that will be used in CNN and BiLSTM are 32 filters, TF-IDF with max features of 10.000, and epoch 5. In this study, use epoch 5 because when using epoch 10, as shown in Figure 6, there was an increase in the loss value from epoch 6 to 10, reaching above 80%, which resulted in lower accuracy. Therefore, this study chose epoch five as the limit to avoid obtaining low accuracy values.

Results in Table 5 show three models, the use of Unigram has a higher accuracy value than Bigram and Trigram of all data split proportions. However, there is an increase in the accuracy value when using Allgram and Unigram+Bigram. For the following scenario, the Baseline used is TF-IDF with Unigram + Bigram, and the proportion of split data is 90:10 because, based on the results in Table 5, it gets the best accuracy value compared to the other combinations. Unigram + Bigram can outperform other n-grams because Bigram can add context that Unigram does not capture and does not include trigrams which can cause many features and overfitting, which can reduce accuracy and results of Table 5 prove that Allgram has decreased compared to Unigram + Bigram.

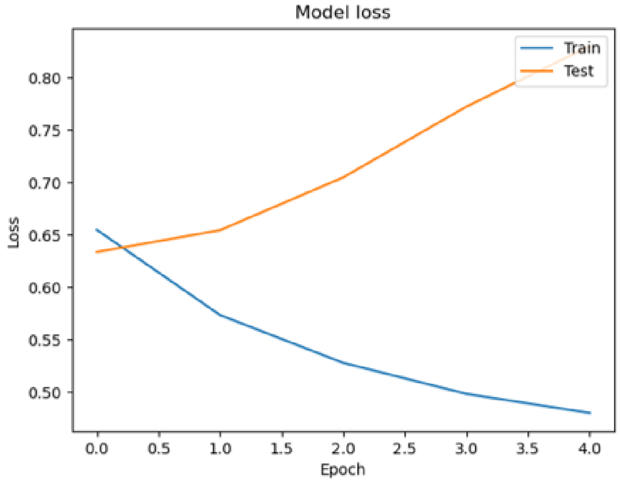

Figure 6. Model loss

Table 5. Result of CNN and BiLSTM scenario

Model	TF-IDF	Accuracy (%)		
		90:10	80:20	70:30
CNN	Unigram	77.43	77.72	77.75
	Bigram	72.95	73.08	72.53
	Trigram	59.75	60.36	60.12
	Allgram	77.94	77.40	77.00
	Unigram+Bigram	**78.69**	77.89	77.25
BiLSTM	Unigram	77.2	77.29	77.07
	Bigram	73.33	72.6	71.67
	Trigram	59.35	59.6	59.19
	Allgram	77.23	77.08	76.76
	Unigram+Bigram	**77.45**	77.15	76.56

3.2 Hybrid Deep Learning CNN-BiLSTM and BiLSTM-CNN

Furthermore, in this scenario, a combination between CNN and BiLSTM will be combined to become a hybrid Model of CNN-BiLSTM and BiLSTM-CNN to detect cyberbullying using the Baseline determined in the previous scenario. The CNN and BiLSTM parameters used are 32 filters and 64 batch sizes with the same epoch as before, epoch 5.

Table 6. Result of Hybrid Deep Learning scenario

Model	Accuracy (%)		
	Baseline	CNN-BiLSTM	BiLSTM-CNN
CNN	78.69	77.47 (-1.22)	-
BiLSTM	77.45	-	77.44 (-0.01)

From the results of the scenarios in Table 6, the CNN-BiLSTM model produces an accuracy value of 77.47% and has decreased in accuracy compared to the CNN model. The BiLSTM-CNN model has an accuracy value of 77.44% and has decreased compared to BiLSTM model. However, we can still improve the accuracy value in the following scenario by using Feature Expansion.

3.3 Feature expansion with FastText

This scenario aims to determine the effect of feature expansion on the Hybrid Deep Learning model and the two models, CNN and BiLSTM, as a comparison. FastText will be used for building a corpus and as a feature expansion by testing some of the most similar Top Rank words in the corpus, including Top 1, Top 5, Top 10, Top 20, Top 35, and Top 40, to find out the best Top Rank. First, it will test on the pre-trained corpus provided by FastText.

Table 7. Result of FastText pre-trained scenario

Model	Top Rank	Accuracy (%)	
		Baseline	Pre-trained
CNN	Top 1	78.69	78.31 (-0,38)
	Top 5		78.64 (-0,05)
	Top 15		79.37 (+0,68)
	Top 20		79.15 (+0,46)
	Top 35		**79.74 (+1,05)**
	Top 40		79.01 (+0,32)
BiLSTM	Top 1	77.45	77.12 (-0,33)
	Top 5		78.95 (+1,5)
	Top 15		79.7 (+2,25)
	Top 20		79.71 (+2,26)
	Top 35		**80.01 (+2,56)**
	Top 40		79.55 (+2,10)
CNN-BiLSTM	Top 1	78.69	77.52 (-1,17)
	Top 5		78.84 (+0,15)
	Top 15		79.32 (+0,63)
	Top 20		79.34 (+0,65)
	Top 35		**80.02 (+1,33)**
	Top 40		79.66 (+0,97)
BiLSTM-CNN	Top 1	77.45	77.21 (-0,23)
	Top 5		78.56 (+1,11)
	Top 15		79.35 (+1,90)
	Top 20		79.64 (+2,19)
	Top 35		**80.28 (+2,83)**
	Top 40		79.66 (+2,21)

The results of the scenarios in Table 7 show that almost all models experience an increase in accuracy values from the baseline. The best accuracy value is in the Top 35 in all models tested. The BiLSTM-CNN model achieves the highest accuracy value of the other models, with an accuracy of 80.28%. The CNN-BiLSTM model also shows good performance with an accuracy of 80.02, while the BiLSTM model achieves an accuracy of 80.01. The CNN model has the lowest accuracy among the four models, with a value of 79.74.

In addition, testing was completed in the Top 40 because the increase in accuracy value was only from Top Rank 1 to 35, and in Top Rank 40, failed to perform well from the baseline with a sign (-) followed by a total score that has dropped from the baseline value. Furthermore, testing will be carried out on several corpora that have been made before, including Indonews, Tweets, and Indonews+Tweets.

From Table 8, the corpus tested obtained the best accuracy value in the Top 1 for each model. The results of the three corpora, Indonews, Tweet, and Indonews+Tweet, have the same pattern, the highest accuracy value for each corpus is in the CNN-BiLSTM model with an accuracy value of Indonews 80%, Tweet 78.95%, and Indonews+Tweet 79.93%. Then, the BiLSTM model with an accuracy value of Indonews 79.73%, Tweet 78.86%, and Indonews+Tweet 79.86% and the BiLSTM-CNN model with an accuracy value of Indonews 79.65%, Tweet 78.77%, and Indonews+Tweet 79.79%. The model with the lowest accuracy score is CNN, with Indonews at 79.28%, Tweet at 78.47%, and Indonews+Tweet at 79.72%. Then there, 19.44% of the total accuracy values obtained in Table 8 failed to perform well from the baseline indicated by the (-) sign, followed by the total value, which decreased from the baseline value, and the rest experienced an increase from the baseline indicated by the (+) sign, followed by a total value that is increased from the baseline value.

Table 8. Result of built corpus scenario

Model	Top Rank	Baseline	Accuracy (%)		
			Indonews	Tweet	Indonews + Tweet
CNN	**Top 1**	78.69	**79.28 (+0.59)**	**78.47 (-0,22)**	**79.72 (+1,03)**
	Top 5		78.87 (+0.18)	78.42 (-0,27)	78.77 (+0,08)
	Top 15		79.28 (+0.59)	76.24 (-2,45)	79.26 (+0,57)
BiLSTM	**Top 1**	77.45	**79.73 (+2.28)**	**78.86 (+1,41)**	**79.86 (+2,41)**
	Top 5		79.66 (+2.21)	78.11 (+0,66)	79.74 (+2,29)
	Top 15		79.22 (+1.77)	76.48 (-0,97)	79.36 (+1,91)
CNN-BiLSTM	**Top 1**	78.69	**80.00 (+1,31)**	**78.95 (+0,26)**	**79.93 (+1,24)**
	Top 5		79.62 (+0.93)	78.67 (-0,02)	79.87 (+1,18)
	Top 15		79.37 (+0,68)	76.01 (-2,68)	79.68 (+0,99)
BiLSTM-CNN	**Top 1**	77.45	**79.65 (+2,20)**	**78.77 (+1,32)**	**79.79 (+2,34)**
	Top 5		79.66 (+2,21)	78.26 (+0,81)	79.72 (+2,27)
	Top 15		79.30 (+1,85)	75.56 (-1,87)	79.34 (+1,89)

3.4 Hypertuning parameter

In this last scenario, hypertuning will be carried out on the existing parameters in the Hybrid Deep Learning, CNN-BiLSTM and BiLSTM-CNN models to increase the accuracy value. Both models were tested with Baseline combined with FastText Top Rank 35 from pre-trained corpus because the previous scenario got the best accuracy value compared to other Top Rank and Corpus. Table 9 presents the parameters that will test.

Table 9. List of parameters

Filter CNN	Dropout	Filter BiLSTM	Dropout	Dense
32	-	32	-	32
32	-	200	-	32
64	-	128	-	32
64	-	128	0.5	32
64	0.5	128	0.5	32
64	0.5	256	0.5	32
64	0.8	256	0.8	32
64	0.5	256	0.5	64
64	0.5	256	0.5	128
64	-	256	-	64
128	0.5	256	0.5	32

Parameter testing on the CNN-BiLSTM and BiLSTM-CNN models built various accuracy values and has an increase in accuracy values on several parameter experiments. The best accuracy value for the CNN-BiLSTM model is 80.55% with the parameters CNN 64 Filter, dropout 0.5, BiLSTM 256 filter, dropout 0.5 and Dense 64. The best accuracy value for the BiLSTM-CNN model is 80.35% with BiLSTM 128 filter parameters, dropout 0.5, CNN 64 Filter, and Dense 32.

3.5 Discussion

Testing in each scenario has the best value for each model. For the first scenario in Table 5, the Baseline obtained the best accuracy values for the CNN model at 78.69% and BiLSTM at 77.45% with a split ratio of 90:10 and TF-IDF Unigram-Bigram. In the following scenario, the CNN-BiLSTM hybrid model experienced a decrease from CNN with an accuracy value of 77.47%, and BiLSTM-CNN also experienced a slight decrease when compared to the CNN model with an accuracy value of 77.44% based on Table 6.

In the third scenario, a combination is performed with FastText for building a corpus and as feature expansion by using several Top Rankings on word similarity to see the effect on the accuracy value. This study conducts tests using two types of the corpus, the pre-trained corpus provided by FastText and the self-built corpus, including Indonews, Tweet, and Indonews+Tweet. Based on Table 7 and Table 8, the corpus which has the best score is corpus pre-trained FastText with Top Rank 35 in each model, CNN at 79.74%, BiLSTM at 80.01%, CNN-BiLSTM at 80.02%, and BiLSTM-CNN at 80.28%. Then, several accuracy values did not perform well from the baseline when combined with Twitter's built corpus. This result could be due to the need for datasets to be used to build the corpus, which only comes from previously collected Twitter datasets. So there still needs to be more vocabulary to do feature expansion.

This scenario proves that the use of FastText as a feature expansion can affect the increase in accuracy, and the combination between hybrid and FastText gets the highest increase in accuracy value with a 1.33% increase in CNN-BiLSTM compared to Baseline CNN. Then, the increase for BiLSTM-CNN is 2.83% compared to Baseline BiLSTM. The increase in accuracy is due to FastText better than just using TF-IDF in representing tweets through word vectors because of the diversity of words in the tweets used [7].

Furthermore, this study does hypertuning parameter to improve the Hybrid Deep Learning model's accuracy, which is the study's aim using the parameters in Table 9. The results obtained for the best accuracy value on CNN-BiLSTM were 80.55%, and there was an increase in accuracy value of 1.86% compared to Baseline CNN. The accuracy value for BiLSTM-CNN is 80.35%, with an increase in accuracy 2.90% compared to Baseline BiLSTM. A summary of the best accuracy values in each scenario is in Figure 7.

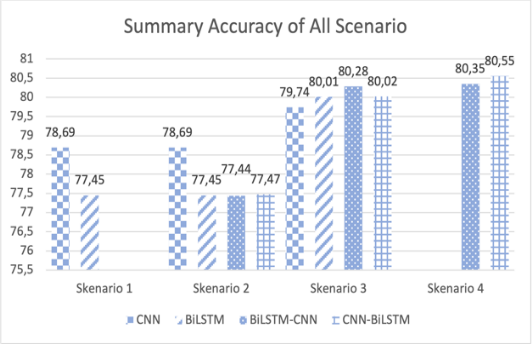

Figure 7. Summary accuracy of all scenario

Table 10. Comparison with related study

Authors	Dataset	Model	Feature Extraction	Feature Expansion	Accuracy
Muzakir et al. [14]	1,065	SVM	Bag of Words	-	76%
Waisnawa	10,000	SVM	TF-IDF	-	76.2%
Laxmi et al. [16]	725	CNN	Doc2Vec	-	65%
Proposed method	29,085	CNN + BiLSTM	TF-IDF	FastText	**80.55%**
Proposed method	29,085	BiLSTM + CNN	TF-IDF	FastText	**80.35%**

Table 10 compares the proposed method created by this study and research related to cyberbullying detection using datasets from Indonesian-language Twitter. The proposed method in this study produces the highest accuracy value compared to other studies. This result shows that the use of CNN and BiLSTM as Hybrid Deep Learning is going quite well by using the capabilities of each model previously described and understanding the context of sentences assisted by FastText as feature expansion to make sentences more meaningful by finding similarity values from various vocabularies provided by the dataset.

4. CONCLUSION

This paper has detected cyberbullying using Hybrid Deep Learning and FastText as feature expansion in Indonesian-language Twitter. This study conducted several tests to achieve the purpose. First, to find the most optimal baseline from several n-grams of TF-IDF. Then, combine to become Hybrid Deep Learning and use FastText as feature expansion. In the last test, this study performed hypertuning on the parameters of the hybrid model, CNN-BiLSTM and BiLSTM-CNN. Based on the test results, the proposed hybrid model with hypertuning combined with FastText as a feature expansion achieved the best accuracy value from all the scenarios done with CNN-BiLSTM, which is 80.55% and BiLSTM-CNN at 80.35%. This result shows that combining Hybrid Deep Learning and FastText as a feature expansion to detect cyberbullying reached a very high increase with an increase value for CNN-BiLSTM of 1.86% and BiLSTM-CNN of 2.90% from the baseline. We hope that the results of this study can be more helpful than previous research with a better accuracy value for detecting cyberbullying on Indonesian-language Twitter because there is FastText as a feature expansion that can handle Indonesian-language sentences with various languages previously explained, such as regional languages, slang, and abbreviations in words. Thus, the results of this study can enable the deletion of tweets containing cyberbullying to be more accurate and on target, fostering a sense of secure for users of the Twitter platform. For further research, test other Hybrid Deep Learning combinations and use other feature expansions to determine the effect of increasing the accuracy value. Then, add variations to the dataset because cyberbullying has many types and variations.

REFERENCES

[1] Sharma, K., Gope, L. (2022). Importance of social media in education sector: A theoretical introspect. International Journal of Pedagogy, Innovation & New Technologies, 9(2): 128-133. https://doi.org/10.5604/01.3001.0016.3222

[2] Yaqub, M. (2022). How Many Tweets per Day 2022 (New Data). https://www.businessdit.com/number-of-tweets-per-day/.

[3] Feinberg, T., Robey, N. (2009). Cyberbullying: Intervention and prevention strategies. National Association of School Psychologists, 38: 1-4. https://wsasp.org/resources/Documents/Mental%20Health/15-1_S4-15.pdf.

[4] Jayani, D.H. (2019). Survei APJII: 49% Pengguna Internet Pernah Dirisak di Medsos. https://databoks.katadata.co.id/datapublish/2019/05/16/survei-apjii-49-pengguna-internet-pernah-dirisak-di-medsos#:~:text=Survei%20Penetrasi%20Internet%20dan%20Perilaku%20Pengguna%20Internet%20di,pengguna%20internet%20yang%20tidak%20pernah%20dirisak%20sebesar%2047%2C2%25.

[5] Nixon, C.L. (2014). Current perspectives: The impact of cyberbullying on adolescent health. Adolescent Health, Medicine and Therapeutics, 143-158. https://doi.org/10.2147/AHMT.S36456

[6] Setiawan, E.B., Widyantoro, D.H., Surendro, K. (2016). Feature expansion using word embedding for tweet topic classification. In 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA), Denpasar, Indonesia, pp. 1-5. https://doi.org/10.1109/TSSA.2016.7871085

[7] Kaibi, I., Satori, H. (2019). A comparative evaluation of

word embeddings techniques for twitter sentiment analysis. In 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, pp. 1-4. https://doi.org/10.1109/WITS.2019.8723864

[8] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5: 135-146. https://doi.org/10.1162/tacl_a_00051

[9] Joshi, R., Gupta, A., Kanvinde, N. (2022). Res-CNN-BiLSTM Network for overcoming mental health disturbances caused due to cyberbullying through social media. arXiv: 2204.09738. https://doi.org/10.48550/arXiv.2204.09738

[10] Aldhyani, T.H., Al-Adhaileh, M.H., Alsubari, S.N. (2022). Cyberbullying identification system based deep learning algorithms. Electronics, 11(20): 3273. https://doi.org/10.3390/electronics11203273

[11] Dewani, A., Memon, M.A., Bhatti, S. (2021). Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data. Journal of Big Data, 8(1): 160. https://doi.org/10.1186/s40537-021-00550-7

[12] Nurrahmi, H., Nurjanah, D. (2018). Indonesian twitter cyberbullying detection using text classification and user credibility. In 2018 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 543-548. https://doi.org/10.1109/ICOIACT.2018.8350758

[13] Andriansyah, M., Akbar, A., Ahwan, A., Gilani, N.A., Nugraha, A.R., Sari, R.N., Senjaya, R. (2017). Cyberbullying comment classification on Indonesian selebgram using support vector machine method. In 2017 Second International Conference on Informatics and Computing (ICIC), Jayapura, Indonesia, pp. 1-5. https://doi.org/10.1109/IAC.2017.8280617

[14] Muzakir, A., Syaputra, H., Panjaitan, F. (2022). A comparative analysis of classification algorithms for cyberbullying crime detection: An experimental study of twitter social media in indonesia. Scientific Journal of Informatics, 9(2): 133-138. https://doi.org/10.15294/sji.v9i2.35149

[15] Putri, N.L.P.M.S., Nurjanah, D., Nurrahmi, H. (2022). Cyberbullying detection on twitter using support vector machine classification method. Building of Informatics, Technology and Science (BITS), 3(4): 661-666. https://doi.org/10.47065/bits.v3i4.1435

[16] Laxmi, S.T., Rismala, R., Nurrahmi, H. (2021). Cyberbullying detection on Indonesian twitter using doc2vec and convolutional neural network. In 2021 9th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, pp. 82-86. https://doi.org/10.1109/ICoICT52021.2021.9527420

[17] Anandarajan, M., Hill, C., Nolan, T., Anandarajan, M., Hill, C., Nolan, T. (2019). Text preprocessing. Practical Text Analytics: Maximizing the Value of Text Data, 45-59. https://doi.org/10.1007/978-3-319-95663-3_4

[18] Qaiser, S., Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1): 25-29. https://doi.org/10.5120/ijca2018917395

[19] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G.S., Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4): 5569-5585. https://doi.org/10.1007/s11042-022-13459-x

[20] Yahya, R.A., Setiawan, E.B. (2022). Feature expansion with fasttext on topic classification using the gradient boosted decision tree on twitter. In 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, pp. 322-327. https://doi.org/10.1109/ICoICT55009.2022.9914896

[21] Yue, W., Li, L. (2020). Sentiment analysis using Word2vec-CNN-BiLSTM classification. In 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), Paris, France, pp. 1-5. https://doi.org/10.1109/SNAMS52053.2020.9336549

[22] Rhanoui, M., Mikram, M., Yousfi, S., Barzali, S. (2019). A CNN-BiLSTM model for document-level sentiment analysis. Machine Learning and Knowledge Extraction, 1(3): 832-847. https://doi.org/10.3390/make1030048

[23] Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., Suieuova, N., Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. International Journal of Advanced Computer Science and Applications, 14(5): 396-406. https://doi.org/10.14569/IJACSA.2023.0140542

[24] Park, J., Kim, C., Dinh, M.C., Park, M. (2022). Design of a condition monitoring system for wind turbines. Energies, 15(2): 464. https://doi.org/10.3390/en15020464