Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 7, No. 1, Maret 2025, Hal. 74-84 e-ISSN 2685-5518

Hate Speech Detection using Expansion Feature Glove with CNN and Bi-LSTM on Twitter

Muchammad Alfi Karom[1*], **Erwin Budi Setiawan**[2],
[1,2] Informatics, School of Computing, Telkom University, Bandung, Indonesia
[*] Email: [1]alfikarom@telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstract

Twitter is one of the biggest social media digital platforms in Indonesia. It serves a medium for readers worldwide and also can be used a means of disseminating information for everyone. However, some people in social media misuse it to spread hate speech against some certain group or communities. Because hate speech it happens everywhere, we need a system to detect hate speech. Sometimes to detect hate speech in Twitter in can be very difficult because lack of context. Needing feature for this problem can make detect hate speech become more easier. Glove is a feature expansion method combine with feature extraction using N-gram and Term Frequency Inverse Document Frequency(TF-IDF) as a method. Data from that it will processed using a hybrid deep learning that combines Convolutional Neural Networks(CNN) dan Bidirectional Long Short-Term Memory(Bi-LSTM). In this study, author obtained 69,484 data related to hate speech. From this study combine feature extraction and feature expansion method has an impact on this research. Best accuracy with all of method is CNN+Bi-LSTM Hybrid method with 91,69% accuracy on top10. Meanwhile best method for Bi-LSTM+CNN method is 91,33% accuracy on top20.

Keywords*: feature expansion, Glove, hate speech, hybrid deep learning, Twitter*

INTRODUCTION

Social media is an medium that allows users to represent themselves and interact in social media , collaborate, share, and communicate with other users, connecting social virtually(Septian, Fachrudin and Nugroho, 2019). One of the popular social media platforms in Indonesia is Twitter. Twitter itself is one of the most widely used platforms, with around 19.5 million users out of 245 million users in Indonesia(Kominfo, 2022). Twitter is a social media platform that has the concept of disseminating information in a concise, direct, and real-time manner, with messages limited to less than 280 characters. Twitter serves as a medium for readers around the world and can be used as a means of spreading information to everyone(Fadli and Hidayatullah, 2021). However, many netizens misuse to use this platform to spread hate speech against certain ethnic groups or communities. According to Law No. 11 of 2008 in Indonesia, there are certain restrictions on how to communicate through social media(Eka Sembodo, Budi Setiawan and Abdurahman Baizal, 2016). Therefore, research about hate speech in advanced is needed to detect hate speech problem on Twitter.

Research on Twitter in Indonesian language has been conducted, in studies(Eka Sembodo, Budi Setiawan and Abdurahman Baizal, 2016; Isnain, Sihabuddin and Suyanto, 2020) data crawling was used to obtain some datasets of tweets in the Indonesian language. Based on previous research, researchers using data crawling to obtain datasets from Twitter in Indonesian language. Hybrid Deep Learning was introducing as a method to detection hate speech in general. In this study(Dewi and Ciptayani, 2022), Hybrid Deep Learning was performed by combining two or more methods to achieve better accuracy for detection hate speech. In another study(Salur and Aydin, 2020) used Hybrid Deep Learning with a CNN+BiLSTM approach, achieving the best accuracy of 82.14% compared to using standard Deep Learning methods, where Convolutional Neural Networks (CNN) had an accuracy of only 78.06%, and Bidirectional Long Short Term Memory (Bi-LSTM) had an accuracy of 80.44%. This study proves that using Hybrid Deep Learning can enhance the accuracy of a method.

The Glove method was using for feature expansion technique at reducing problem about discrepancies of words on Twitter in the Indonesian language. This is necessary because tweets on Twitter especially in Indonesian are often difficult to understand without context. Therefore, a feature is needed to facilitate this understanding. In study(Lim, Setiawan and Santoso, 2019), Glove outperformed Word2Vec, achieving an average F1 accuracy of 0.476 compared to Word2Vec's F1

score of 0.470. this researcher chose to use Glove for feature expansion due to its limited use in combination with Hybrid Deep Learning.

Study(Carracedo and Mondéjar, 2021), inform hate speech detection using Bidirectional Long Short Term Memory achieved a better accuracy of 79% compared to using SVM, which only reached 75%. However, that research only focused on the CNN method. Therefore, the researcher intends to utilize both CNN and Bi-LSTM methods, as CNN shown the best accuracy among all methods. Meanwhile, the Bi-LSTM method has not been previously applied in the Indonesian language. This research is expected to improve the performance of feature expansion in detecting hate speech. Also Study (D'Sa, Illina and Fohr, 2020) With Combined two different metohods CNN and Bi-LSTM been conducted with different types of embeddings, specifically FastText and BERT. The result from classification model using CNN showed an accuracy of approximately 91.5% with FastText and 90.9% with BERT. Bi-LSTM method achieved an accuracy of 91.9% with both FastText and BERT. the combined method produced an accuracy of 97.0%, indicating with the combine method has potential to yield better for classification performance. In this study also employs methods different from previous research. Besides using CNN and Bi-LSTM also combined with feature extraction and feature expansion with the data used sourced from Twitter has been processed through stop words removal. Data from Twitter also divded to several categories.

RESEARCH METHOD

The data obtained from this research comes from the social media platform Twitter. The collected data will be labeled and then followed by a preprocessing method that utilizes feature extraction through N-gram combined with Glove. Glove itself is an unsupervised learning method for words that surpasses models such as word analogy, word similarity, and recognition(Ali *et al.*, 2022). The results will be divided into training data and testing data, with the testing data used as a tool for classification trials using various methods, after which the model will produce performance results from the tests.

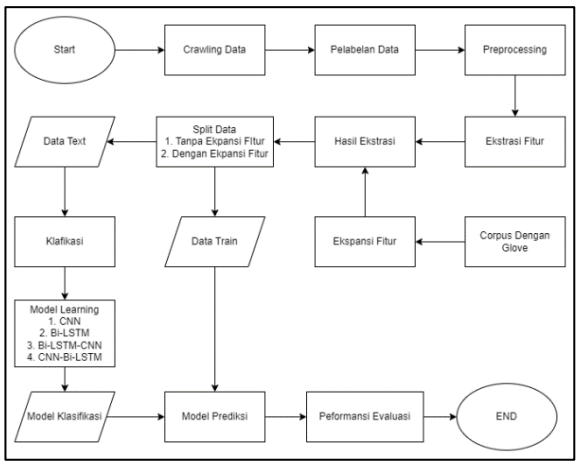

Figure 1. Hate Speech Detection System

2.1 Data Acquisition

The data obtained from Twitter will lead data crawling method obtained through the Application Program Interface (API) in the Indonesian language. The Indonesian data will then be separated into several topics, as shown in Table 1.

Table 1. Hate Speech Topic

Topic	Amount
Agama	13,909
Politik	10,059
Ras	2,500
Kepolisian	10,003
Orientasi Seksual	10,150
Covid-19	10,034
Kata Kasar	12,809
Total	**69,484**

2.2 Data Labeling

Data labeling will be performed on the dataset before the classification process. The researcher will label each word that goes through the classification process(Khan *et al.*, 2022). This process will yield labels consisting of Hate Speech (HS), which refers to sentences that containing elements of hate speech, and Non-Hate Speech (NHS), which refers to sentences that do not contain elements of hate speech. An example of labeling is shown in the following Table 2.

Table 2. Data Labeling

Sentence	Classification
INGAT……..PADORU Budaya Kebaratan ANjiing	HS
Pemerintah HARUSNYA membaguskan Jalan DI Indo	NHS
Hadeh… Admin Jembut but but mana kau???	HS

2.3 Preprocessing

The Twitter data will undergo with filtering process to eliminate non-important aspects in the Twitter sentences. The word in Indonesian Twitter data will subsequently be cleaned using the applied methods.

2.3.1 Data Cleansing

Data Cleaning process, the data we obtained from Twitter will be filtered to remove noise such as emoticons, numbers, and punctuation marks (Septian, Fachrudin and Nugroho, 2019). As seen in Table 3.

Table 3. Data Cleansing

Before	After
INGAT…… PADORU Budaya di Kebaratan Anjiing	INGAT PADORU Budaya di Kebaratan Anjiing
Pemerintah HARUSNYA membaguskan Jalan DI Indo	Pemerintah HARUSNYA membaguskan Jalan DI Indo
Hadeh… Admin Jembut but but mana kau???	Hadeh Admin Jembut but but mana Kau

2.3.2 Stop Words

Stop words are the process of removing common words that are not significant in the classification process. The researcher will create a list of stop words in Indonesian, which will be eliminated from the sentences(Septian, Fachrudin and Nugroho, 2019). As seen in Table 4.

Table 4. Stop Words

Before	After
INGAT PADORU Budaya di Kebaratan Anjiing	INGAT PADORU Budaya Kebaratan Anjiing
Pemerintah HARUSNYA membaguskan Jalan DI Indo	Pemerintah HARUSNYA membaguskan Jalan Indo
Hadeh Admin Jembut but but mana Kau	Hadeh Admin Jembut but but mana

2.3.3 Stemming

In Stemming, words with affixes will be transformed into their root forms by removing the affixes (Prihatini, 2016). As seen in Table 5.

Table 5. Stemming

Before	After
INGAT PADORU Budaya Kebaratan Anjiing	INGAT PADORU Budaya barat Anjiing
Pemerintah HARUSNYA membaguskan Jalan Indo	Pemerintah HARUS bagus Jalan Indo
Hadeh Admin Jembut but but mana	Hadeh Admin Jembut but but mana

2.3.4 Case Folding

Case Folding is the stage where sentences in Twitter that contain uppercase letters are changed to lowercase (Prihatini, 2016). As seen in Table 6.

Table 6. Case Folding

Before	After
INGAT PADORU Budaya barat Anjiing	ingat padoru budaya barat anjiing
Pemerintah HARUS bagus Jalan Indo	pemerintah harus bagus jalan indo
Hadeh Admin Jembut but but mana	hadeh admin Jembut but but mana

2.3.5 Tokenizing

Tokenizing is the stage of separating words in a sentence into individual words that are independent, using a space (‘’) as a delimiter (Septian, Fachrudin and Nugroho, 2019). As seen in Table 7.

Table 7. Tokenizing

Before	After
ingat padoru budaya barat anjiing	‘ingat’ ‘padoru’ ‘budaya’ ‘barat’ ‘anjiing’
pemerintah harus bagus jalan indo	‘pemerintah’ ‘harus’ ‘bagus’ ‘jalan’ ‘indo’
hadeh admin Jembut but but mana	‘hadeh’ ‘admin’ ‘jembut’ ‘but’ ‘mana’

2.4 Feature Extraction

Feature extraction is used to obtain the words contained in the feature documents present in a tweet, assigning values to those word(Yu *et al.*, 2018). In the feature extraction stage. The tweet representation in this research uses Boolean features, where each feature indicates the presence or absence of a word in the tweet.

2.5 Corpus

In this corpus, Glove is used as a method, and the corpus will align with the rank of similarity to an individual word(Oprea and Magdy, 2020). The corpus with the highest rank will be used for feature expansion. This way, it can reduce the workload on the computer, thereby limiting the rank of a word and constraining its usage. In the feature extraction stage, N-gram will be performed as explained; the researcher wants to take an example from the sentence for unigram, and thus the feature extraction will appear as follows:

	ingat	padoru	budaya	barat	anjiing	pemerintah	harus	bagus	jalan	indo	hadeh	admin	jembut	but	mana
doc 1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
doc 2	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
doc 3	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Figure 2. N-Gram Visualization

After conducting N-gram feature extraction research, the data that has been tokenized will undergo another feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF) to obtain the weight of a particular word. The process of TF-IDF will be as shown in Table 8.

Table 8. Bad Words with TF-IDF

No	Word	TF-IDF
1	anjiing	0.533333
2	jembut	0.533333
3	hadeh	0.266667
4	pemerintah	0.066667
5	admin	0.066667
6	padoru	0.003182
7	harus	0.003182
8	indo	0.002120
9	jalan	0.002120
10	bagus	0.000000

2.6 Feature Expansion

Feature expansion in this research is used to identify missing words in the tweet representation, then substitute them with semantically related words. This method is primarily used to complement and find the missing words that have similarities, which functions to achieve optimal output (Siregar, 2022).

2.7 CNN

Convolutional Neural Networks (CNN) are one of the popular classifiers for stance classification. CNN was first used by Collobert to automatically extract important features (Widayati, 2018). This hierarchical feature information is obtained through max-over-time pooling operations, further developed by Kim for sentence classification. CNN is also utilized to perform a method by filtering words through computation. This can be seen in Figure 3.

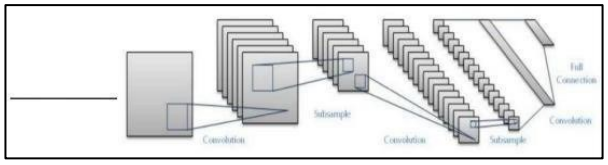

Figure 3. CNN Visualization

2.8 Bi-LSTM

As shown in Figure 3, Bidirectional Long Short-Term Memory (Bi-LSTM) is a class of RNN model that addresses the vanishing gradient problem. Bi-LSTM is used for sequential data processing and is efficient in capturing long-range dependencies(D'Sa, Illina and Fohr, 2020). Thus, Bidirectional LSTM involves two layers that process information in opposite directions. This model is very effective for recognizing patterns in sentences since each word in a document is processed sequentially, allowing tweets to be understood when learned in the order of each word. The lower layer moves forward, processing from the first word to the last, while the upper layer moves backward, processing from the last word to the first. Bi-LSTM is particularly beneficial for sequential labeling as it has access to information from both before and after the current word (Fadli and Hidayatullah, 2021).

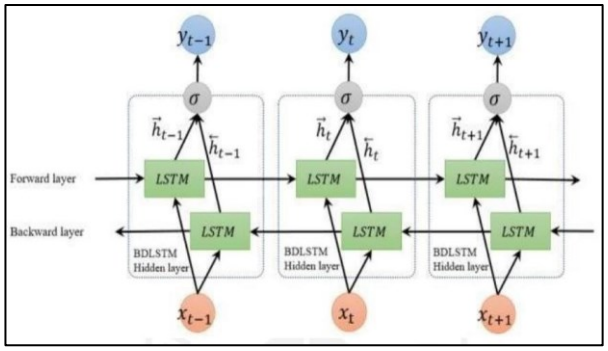

Figure 4. Bi-LSTM Visualization

2.9 Hybrid Model

This initial method is also referred to as Hybridization(Sun, Wang and Tang, 2013). The

Hybrid Model combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) methods aimed at achieving optimal results. In the figure 5, the initial method uses CNN as the starting model, which is then combined with the Bi-LSTM method. In the subsequent experiment, this method is reversed, starting with Bi-LSTM followed by the CNN method. The scheme is illustrated in the image below.

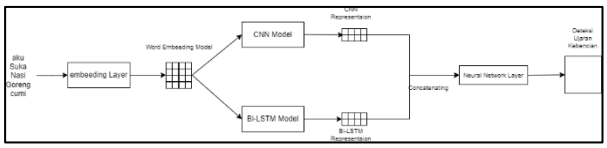

Figure 5. Hybrid Model CNN+Bi-LSTM

2.10 Evaluation

Evaluation is used in Machine Learning to monitor or determine the behavior of a classification model. The structure of this evaluation is presented in rows and columns, where rows represent the actual classes of instances and columns represent the predicted classes(Hasnain *et al.*, 2020). This evaluation results in metrics such as Accuracy, Precision, F1-score, and Recall. The Confusion Matrix is represented in a 2x2 format comprising four measurement forms: "True Positive," "True Negative," "False Positive," and "False Negative." From the frequency of these four components, indicators of the classifier's performance in detecting the given class can be obtained by calculating accuracy, precision, and recall based on the constructed algorithm(Prabowo *et al.*, 2021).

Table 9. Confusion Matrix

Correct Clasification	**Classfied AS**	
	+	-
+	True Positif (TP)	False Negatif (FN)
-	False Positif (FP)	True Negatif (TN)

Precision measures the proportion of the positive class that was correctly predicted out of the total positive class, while Recall indicates the percentage of positive data class that was correctly predicted from the entire positive data class. On the other hand, Accuracy is the ratio of correct predictions to the total data. It can be calculated using the formula:

1. Precison = TP/(TP+FP)
2. Recall = TP/(TP+FN)
3. Accuracy =(TP+TN)/(TP+TN+FP+FN)

RESULT AND DISCCUSION

in this research, we explain how a method works as well as the purpose of that method. This experiment involves four methods: CNN, Bi-LSTM, Bi-LSTM+CNN, and CNN+Bi-LSTM.

3.1 Testing and Result

in the first experiment, the method involves incorporating the TF-IDF method into feature extraction to find the best data split and determine the maximum features. For the CNN method, the parameters used in this experiment include a batch size of 64, 32 filters, a dropout of 0.5, and 10 epochs. For the Bi-LSTM method, the parameters also include a batch size of 64, 64 filters, a dropout of 0.5, and 10 epochs. Initially, the unigram method was used. This experiment also involved using maximum and minimum features divided into 5000, 10000, and 15000 to see if features have a significant impact on the experiment. Additionally, a word splitting technique was employed where the words created would be randomized by the system, and then the data would be split based on the number of sentences in the text. In this experiment, we used three methods to distribute the number of words into three types: 90:10, where the training data consists of 90% and 10% of the remaining data is used as the test data; 80:20, meaning 80% for training and 20% for testing; and 70:30, meaning 70% for training and 30% for testing. We also compared the maximum features to see if having more reference data could be beneficial.

Based on the table 10, it can be seen that for the CNN method, the best performance was achieved with a data split ratio of 80:20 and a maximum of 10000 features, resulting in an accuracy of 90.11%. For the Bi-LSTM method, the data split of 90:10 with a maximum of 10000 features achieved an accuracy of 90.90%, making it the highest accuracy method compared to others. For the Bi-LSTM+CNN method, the best results were with a data split of 90:10 and a maximum of 5000 features, achieving an accuracy of 90.83%. For the CNN+Bi-LSTM method, the best performance was with a data split of 80:20 and a maximum of 5000 features, resulting in an accuracy of 90.20%. The accuracy results were not too far apart, indicating that we avoided overfitting the data during this experiment.

Table 10. Data Size and Max-Features Calculation

Max_Features	Data Size	Accuracy (%)			
		CNN	BI-LSTM	Bi-LSTM+CNN	CNN+BI-LSTM
5000	90 : 10	90,02	**90,90**	**90,83**	89,70
	80 : 20	89,81	90,02	90,33	**90,20**
	70 : 30	89,87	90,58	90,28	89,70
10000	90 : 10	89,90	90,11	90,15	88,98
	80 : 20	**90,11**	89,97	89,83	88,85
	70 : 30	89,65	89,66	89,73	89.06
15000	90 : 10	90,00	90,11	90,15	88,98
	80 : 20	89,29	89,68	89,63	89,45
	70 : 30	89,36	89,58	89,40	89,17

The results from the previous experiment led to the establishment of a baseline model, which was then tested against various N-grams. The N-gram testing involved several types, including unigram, bigram, trigram, unigram-bigram, and unigram-bigram-trigram. In this second experiment, the CNN method showed that the Unigram-Bigram approach had the highest accuracy of 90.35%. In the Bi-LSTM method, the unigram approach had the highest accuracy at 90.90%, making it the top-performing method among the trials. For the Bi-LSTM+CNN method, the unigram also achieved the highest accuracy of 90.83%. In the CNN+Bi-LSTM method, the unigram achieved an accuracy of 90.20%, making it the best among the other methods. The results of these methods can be seen in the table.

Table 11. N-Gram Calculation

N-Gram	Accuracy (%)			
	CNN	BI-LSTM	Bi-LSTM+CNN	CNN+BI-LSTM
Baseline	90,11	**90,90**	**90,83**	**90,20**
Bigram	80,18	81,86	82,37	81,33
Trigram	72,95	73,54	73,92	74,45
Unigram-Bigram	**90,35**	89,99	90,57	89,43
Unigram-Bigram-Trigram	90,08	90,39	90,55	89,93

In the third experiment, we tested feature expansion based on the previously established baseline using N-grams. This experiment aimed to find the best model created with the Glove corpus. The objective was to determine the words most similar to the corpus, providing an optimal measurement of similarity with the best accuracy. In this corpus, we conducted tests on Twitter, searching for similarity based on the top 1, top 5, top 10, top 15, and top 20 words. In the CNN method, the Top 1 similarity achieved the best accuracy of 90.70%, while the Bi-LSTM baseline model maintained the highest accuracy of 90.90%. The

Bi-LSTM+CNN model reached its peak accuracy at the Top 20 level with 91.33%. The CNN+Bi-LSTM method exhibited the best accuracy among all the tested methods on Twitter, achieving an accuracy of 91.69% at Top 10. These results can be found in Table 12.

Table 12. Expansion Feature with Twitter Calculation

Tweet	**Accuracy (%)**			
	CNN	**BI-LSTM**	**Bi-LSTM+CNN**	**CNN+BI-LSTM**
Baseline	90,11	**90,90**	90,83	90,20
Top1	**90,70**	90,59	91,01	90,89
	(+0,55%)	(-0,36%)	(+0,14%)	(+0,60%)
Top5	90,07	90,44	90,84	91,45
	(-0,08%)	(-0,54%)	(-0,02%)	(+1,12%)
Top10	90,59	90,71	91,23	**91,69**
	(+0,41%)	(-0,24%)	(+0,35%)	**(+1,36%)**
Top15	90,00	90,42	90,62	90,61
	(-0,15%)	(-0,57%)	(-0,19%)	(+0,36%)
Top20	90,63	90,38	**91,33**	90,57
	(+0,47%)	(-0,60%)	**(+0,44%)**	(+0,33%)

In this experiment, we compared the news data still using the same baseline with the same goal: to find the top 'N' values with optimal accuracy using Glove on news texts. The word similarity tests sought to find the best model from the top 1, top 5, top 10, top 15, and top 20 words. As shown in Table 13, the CNN baseline model achieved the highest accuracy, and using Glove demonstrated a performance close to the baseline but did not surpass it. Similarly, the Bi-LSTM baseline maintained the best accuracy compared to feature expansions. The Bi-LSTM+CNN method still showed the best accuracy at 90.83%. In the CNN+Bi-LSTM method, the accuracy at Top 10 surpassed the baseline, achieving the highest accuracy compared to other methods at 91.03%, as shown in Table 13.

Table 13. Expansion Feature with News Calculation

News	**Accuracy (%)**			
	CNN	**BI-LSTM**	**Bi-LSTM+CNN**	**CNN+BI-LSTM**
Baseline	**90,11**	**90,90**	**90,83**	90,20
Top1	89,95	90,63	90,10	90,23
	(-0,25%)	(-0,31%)	(-0,79%)	(-0,02%)
Top5	88,87	89,80	90,01	**91,09**
	(-1,32%)	(-1,17%)	(-0,97%)	**(0,81%)**
Top10	90,07	90,00	89,14	89,70
	(-0,09%)	(-0,96%)	(-1,82%)	(-0,62%)
Top15	89,60	89,52	90,06	89,71
	(-0,57%)	(-1,51%)	(-0,78%)	(-0,60%)
Top20	89,16	89,17	89,09	89,96
	(-1,04%)	(-1,84%)	(-1,92%)	(-0,29%)

This experiment compares the baseline accuracy levels with the Glove method to find the best 'N' values from two methods based on Twitter and news data, aiming to identify the most optimal top-ranking method with the assistance of Glove. The testing methodology remained the same as previously conducted to find the top models for Top 1, Top 5, Top 10, Top 15, and Top 20. In the CNN method, the Top 1 testing achieved the highest accuracy of 90.84%. In the Bi-LSTM method, the baseline maintained the highest accuracy at 90.90%. For the Bi-LSTM+CNN method, Top 1 achieved an accuracy of 90.90%, representing the highest accuracy among all methods tested in the Twitter and news data experiment. The Bi-LSTM baseline model retained the best accuracy among the top methods. The results of these methods can be seen in Table 14.

Table 14. Expansion Feature with Tweet + News Calculation

Tweet-News	**Accuracy (%)**			
	CNN	**BI-LSTM**	**Bi-LSTM+CNN**	**CNN+BI-LSTM**
Baseline	90,11	**90,90**	90,83	**90,20**
Top1	**90,84**	90,38	**90,90**	89,98
	(0,68%)	(-0,58%)	**(0,02%)**	(-0,28)
Top5	88,72	88,81	89,78	90,00
	(-1,61%)	(-2,25%)	(-1,08%)	(-0,24%)
Top10	89,67	88,89	89,98	89,69
	(-0,53%)	(-2,18%)	(-1,01%)	(-0,56%)
Top15	89,57	88,94	88,58	89,57
	(-0,65%)	(-2,14%)	(-2,48%)	(-0,70%)
Top20	88,72	88,80	89,18	89,11
	(-1,61)	(-2,25%)	(-1,77%)	(-1,20%)

3.2 Discussion

In this search, experiments were conducted through scenario testing to find the best baseline model for a method. The model combines N-gram and Glove corpus models. The developed model was then applied to the best corpus and top-ranked data. Based on the results from the current testing, it is evident that the best model for data splitting is 80:20 for Bi-LSTM and Bi-LSTM+CNN, and 90:10 for CNN and CNN+Bi-LSTM. The optimal maximum feature usage is 5000 for Bi-LSTM, Bi-LSTM+CNN, and CNN+Bi-LSTM, while CNN's optimal maximum feature usage is 10000. The CNN achieved an accuracy of 90.11%, Bi-LSTM 90.20%, Bi-LSTM+CNN 90.83%, and CNN+Bi-LSTM 90.20%.

In the second scenario, tests were conducted based on N-grams, with the Unigram method remaining the best model except for the CNN method, where the Unigram-Bigram method performed best with an accuracy of 90.35%, an increase of about 0.24% from the previous method. In the third testing, conducted using the Glove method on Twitter, the Top 1 method for CNN showed the best performance with an accuracy of 90.70%, an increase of approximately 0.55% over the CNN baseline. The Bi-LSTM baseline method still remains the best method. The Bi-LSTM+CNN achieved an accuracy of 91.33% at Top 20, an increase of about 0.44% from the baseline. The CNN+Bi-LSTM at Top 10 reached an accuracy of 91.69%.

we retested the methods using Indonesian news data. As shown, all baseline methods exhibited the best accuracy for CNN, Bi-LSTM, and CNN+Bi-LSTM, with only the CNN+Bi-LSTM method achieving a higher accuracy of 91.09% at Top 5, an increase of 1.35% from the CNN+Bi-LSTM baseline. In the fifth method, we combined the Twitter and news methods in Indonesian and retested using feature expansion with Glove. The best method for CNN was 90.84% at Top 1, showing an improvement.

The Bi-LSTM baseline remained the best method. The Bi-LSTM+CNN Top 1 method also achieved 90.90%, an increase of about 0.02% from its baseline. Meanwhile, CNN+Bi-LSTM still retained the best baseline method. Based on the model created using Glove, the CNN method for Twitter and news at Top 1 is the best with 90.84%, Bi-LSTM using the baseline with 90.90%, Bi-LSTM+CNN at Top 20 with 91.33%, and CNN+Bi-LSTM using Twitter at Top 10 with 91.69%.

Additionally, a statistical accuracy test was used to validate the accuracy among the results of each scenario created. The concepts of P-Value and Z-Value were applied to enhance the accuracy of the tests. The accuracy indicated in the scenario texts can be considered improved if P-Value < 0.05 and Z-Value > 1.96. As seen in Table 15, the accuracy improvement can be observed from Scenario 1 (Baseline) to Scenario 3 (Corpus).

Tabel 15. Z-Value and P-Value

Model	Parameter	S1-S2	S2-S3	S1-S3
CNN	Z-Value	2.06	5.10	4.85
	P-Value	0.039	0	0.001
	Significant?	**True**	**True**	**True**
Bi-LSTM	Z-Value	0.5	0.5	0.5
	P-Value	0	0	0
	Significant?	**False**	**False**	**False**
Bi-LSTM+CNN	Z-Value	0.5	3.870	3.878
	P-Value	0	0.0001	0.0001
	Significant?	**False**	**True**	**True**
CNN+Bi-LSTM	Z-Value	0.5	14.65	14.71
	p-Value	0	0.0001	0.0001
	Significant?	**False**	**True**	**True**

The results presented in the table demonstrate that using hybrid deep learning can enhance the accuracy of a method. The outcomes of the various methods confirm that using GloVe can reduce the discrepancies in words occurring in Twitter and news data. The words used in Twitter and news can be more varied and shorter, which can be applied effectively. Moreover, combining several words has proven to enhance the benefits of the model in hybrid deep learning compared to the baseline model created. However, the use of these methods may also limit the complexity of certain words due to the integrity of word usage in the model. The accuracy of an experiment will also become more focused on a specific aspect.

CONCLUSION

The detection of hate speech in Indonesian Twitter sentences was carried out using a system developed with four methodological models: CNN, Bi-LSTM, Bi-LSTM+CNN, and CNN+Bi-LSTM. This development utilized feature extraction along with feature expansion. In this feature extraction, TF-IDF was used for calculation word sentences and for feature expansion, Glove was employed to find the similarity of a word with other words that have the same or similar meanings. The best accuracy from all methods was achieved by the hybrid CNN+Bi-LSTM, which reached an accuracy of 91.69% in the Top 10 method. The Bi-LSTM+CNN achieved 91.33% in the Top 20. These results demonstrate that using hybrid deep learning and Glove has a significant impact on accuracy, as evidenced by the increased accuracy and the ability to address issues in detecting words in Indonesian, where few studies have been conducted. The results also confirm that combining several methods such as TF-IDF, Glove, maximum features, and N-Gram plays a significant role to improving the accuracy of a method detection hate speech. This is crucial concerning the detection of hate speech, which can be conducted using local language methods. The hope is that the learning methods created can serve as a reference to enhance the ability to detect hate speech on Twitter or other social media platforms. In the

future, exploring other methods such as RNN, GRU, or alternative techniques may be attempted to further improve method accuracy.

REFERENCES

Ali, R. *et al.* (2022) 'Hate speech detection on Twitter using transfer learning', *Computer Speech and Language*, 74(July), p. 101365. Available at: https://doi.org/10.1016/j.csl.2022.101365.

Carracedo, À.A. and Mondéjar, R.J. (2021) 'Profiling Hate Speech Spreaders on Twitter', *CEUR Workshop Proceedings*, 2936, pp. 1801–1807.

D'Sa, A.G., Illina, I. and Fohr, D. (2020) 'BERT and fastText Embeddings for Automatic Detection of Toxic Speech', *Proceedings of 2020 International Multi-Conference on: Organization of Knowledge and Advanced Technologies, OCTA 2020* [Preprint]. Available at: https://doi.org/10.1109/OCTA49274.2020.9151853.

Dewi, K.C. and Ciptayani, P.I. (2022) 'Pemodelan Sistem Rekomendasi Cerdas Menggunakan Hybrid Deep Learning', *Jurnal Sistem Informasi dan Sains Teknologi*, 4(2), pp. 1–7. Available at: https://doi.org/10.31326/sistek.v4i2.1157.

Eka Sembodo, J., Budi Setiawan, E. and Abdurahman Baizal, Z. (2016) 'Data Crawling Otomatis pada Twitter', (September), pp. 11–16. Available at: https://doi.org/10.21108/indosc.2016.111.

Fadli, H. and Hidayatullah, A. (2021) 'Identifikasi Cyberbullying pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM', *Universitas Islam Indonesia (UII)*, 2(No. 1), pp. 1–6. Available at: https://journal.uii.ac.id/AUTOMATA/article/view/17364.

Hasnain, M. *et al.* (2020) 'Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking', *IEEE Access*, 8, pp. 90847–90861. Available at: https://doi.org/10.1109/ACCESS.2020.2994222.

Isnain, A.R., Sihabuddin, A. and Suyanto, Y. (2020) 'Bidirectional Long Short Term Memory Method and Word2vec Extraction Approach for Hate Speech Detection', *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 14(2), p. 169. Available at: https://doi.org/10.22146/ijccs.51743.

Khan, S. *et al.* (2022) 'BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection', *Journal of King Saud University - Computer and Information Sciences*, 34(7), pp. 4335–4344. Available at: https://doi.org/10.1016/j.jksuci.2022.05.006.

Kominfo (2022) 'Indonesia Peringkat Lima Pengguna Twitter'. Available at: https://www.kominfo.go.id/content/detail/2366/indonesia-peringkat-lima-pengguna-twitter/0/sorotan_media.

Lim, E., Setiawan, E.I. and Santoso, J. (2019) 'Stance Classification Post Kesehatan di Media Sosial Dengan FastText Embedding dan Deep Learning', *Journal of Intelligent System and Computation*, 1(2), pp. 65–73. Available at: https://doi.org/10.52985/insyst.v1i2.86.

Oprea, S. and Magdy, W. (2020) 'iSarcasm: A Dataset of Intended Sarcasm', pp. 1279–1289. Available at: https://doi.org/10.18653/v1/2020.acl-main.118.

Prabowo, C. *et al.* (2021) 'Teknik Klasifikasi Pembayaran SPP Berdasarkan Tingkat Ketepatan Pembayaran', *Jurnal Data Science & Informatika*, 1(1), pp. 1–5.

Prihatini, P.M. (2016) 'Implementasi Ekstraksi Fitur Pada Pengolahan Dokumen Berbahasa Indonesia', *Jurnal Matrix*, 6(3), pp. 174–178.

Salur, M.U. and Aydin, I. (2020) 'A Novel Hybrid Deep Learning Model for Sentiment Classification', *IEEE Access*, 8, pp. 58080–58093. Available at: https://doi.org/10.1109/ACCESS.2020.

2982538.

Septian, J.A., Fachrudin, T.M. and Nugroho, A. (2019) 'Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor', *Journal of Intelligent System and Computation*, 1(1), pp. 43–49. Available at: https://doi.org/10.52985/insyst.v1i1.36.

Siregar, H. (2022) 'Analisis Pemanfaatan Media Sosial Sebagai Sarana Sosialisasi Pancasila', *Pancasila: Jurnal Keindonesiaan*, (1), pp. 71–82. Available at: https://doi.org/10.52738/pjk.v2i1.102.

Sun, Y., Wang, X. and Tang, X. (2013) 'Hybrid deep learning for face verification', *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1489–1496. Available at: https://doi.org/10.1109/ICCV.2013.188.

Widayati, L.S. (2018) 'Ujaran Kebencian: Batasan Pengertian Dan Larangannya', *Info Singkat: Kajian Singkat Terhadap Isu Aktual dan Strategis*, 10(6), pp. 1–6.

Yu, Y. *et al.* (2018) 'A parallel feature expansion classification model with feature-based attention mechanism', *Proceedings of 2018 IEEE 7th Data Driven Control and Learning Systems Conference, DDCLS 2018*, pp. 362–367. Available at: https://doi.org/10.1109/DDCLS.2018.8516066.