Eduvest - Journal of Universal Studies
Volume 4 Number 11, November, 2024
p- ISSN 2775-3735- e-ISSN 2775-3727

HATE COMMENT DETECTION ON TWITTER USING LONG SHORT TERM MEMORY (LSTM) WITH GENETIC ALGORITHM (GA)

Dea Alfatihah Nindya Erlani[1], Erwin Budi Setiawan[2]
[1,2] Fakultas Informatika, Telkom University, Indonesia
Email: deaalfatihah@student.telkomuniversity.ac.id,
erwinbudisetiawan@telkomuniversity.ac.id

ABSTRACT

In the era of social media like today, one social media that is currently quite popular is Twitter. This study explores the use of the Long Short Term Memory (LSTM) method optimized with the Genetic Algorithm (GA) to detect hate speech in Twitter data in Indonesia. We use TF-IDF and GloVe feature extraction techniques to produce effective word vector representations in natural language processing. This study also introduces feature expansion and similarity corpus construction to improve the performance of the LSTM classification model. Evaluation is carried out through a confusion matrix to measure accuracy, precision, recall, and F1 score. The results show that the LSTM model with TF-IDF and GloVe feature extraction achieves the best performance with an accuracy of up to 92.91%. We also found that the combination of Unigram + Bigram + Trigram, max feature 10000, and Glove corpus with Top 20 similarity gave optimal results. In addition, parameter optimization using genetic algorithms has been shown to improve accuracy and F1-Score. The resulting LSTM model is able to classify test data with high accuracy, which has the potential to help in the detection and handling of hate speech on social media, as well as improving the model's ability to identify and understand text content in the Indonesian language context.

KEYWORDS *Hate speech, Twitter, Long Short Term Memory (LSTM), Genetic Algorithm (GA), TF-IDF, GloVe, Detection.*

INTRODUCTION

In the era of social media like today, one social media that is currently quite popular is Twitter. Twitter has become one of the most popular platforms for sharing information, opinions, and interactions between users. However, the increasing use of Twitter also has negative impacts, including the spread of hate speech. Hate speech can include content that insults, threatens, or provokes others based on race, religion, gender, or other factors. The negative impacts of this hate

How to cite: Dea Alfatihah Nindya Erlani, Erwin Budi Setiawan (2024). Hate Comment Detection On Twitter Using Long Short Term Memory (LSTM) With Genetic Algorithm (GA). *Journal Eduvest. 4* (11): 10191-10201
E-ISSN: 2775-3727

speech can include persecution, harassment, or even acts of violence in the real world (Gautam, 2021). With the many cases of hate speech on Twitter, this study will conduct Hate Speech Detection on Twitter. Detection is important because Twitter is one of the most widely used social media platforms for interacting and sharing information. However, along with the development of technology, Twitter has also become a place full of hate speech. Hate speech on Twitter can trigger conflict and affect the mental health of users. Therefore, hate speech detection on Twitter needs to be done. One method that can be used to detect hate speech on Twitter is by using Long Short-Term Memory (LSTM) and Genetic Algorithm optimization. LSTM is a type of artificial neural network that can process data in a certain order, such as text. While Genetic Algorithm is an optimization technique that can be used to find the best parameters in the LSTM model.

However, to improve the accuracy of hate speech detection on Twitter, feature extraction can be done using the TF-IDF method and Glove expansion. TF-IDF is a feature extraction method used to measure how important a word is in a document. While Glove expansion is a technique for enriching word representation by combining word vectors from a larger corpus (Sameer, 2022).

Several studies have been conducted to detect hate speech on Twitter using this method, such as in studies (Bouktif et al., 2018; Pennington et al., 2014; Talita & Wiguna, 2019). these studies show that the LSTM method and Genetic Algorithm optimization with TF-IDF extraction and Glove expansion can improve the accuracy of hate speech detection on Twitter.

The use of deep learning methods such as LSTM also has advantages in overcoming the problem of hate speech detection on Twitter. LSTM is able to process data in a certain order so that it can recognize the context of a sentence or text. This is very important in hate speech detection because hate speech often consists of more than one word or phrase, but also involves the context of the sentence or text as a whole (Chui et al., 2021). In addition, the use of Glove expansion can also improve the accuracy of hate speech detection on Twitter because this technique enriches word representation by combining word vectors from a larger corpus. In research, experimental results showed that the use of Glove expansion can improve the accuracy of hate speech detection by 2-3%.

However, it should be noted that hate speech detection on Twitter is not easy because hate speech is often delivered in ambiguous or unclear language. Therefore, further research is needed to improve the accuracy of hate speech detection on Twitter (Onan, 2021).

In this study, the author proposes to develop a more accurate and efficient hate speech detection model by utilizing a combination of Long Short Term Memory (LSTM) and Genetic Algorithm (GA) methods. By optimizing model parameters using GA and applying feature extraction techniques such as TF-IDF and GloVe, it is expected to improve the model's ability to recognize the context and nuances of language that are often ambiguous in hate speech on social media, especially on the Twitter platform (Tanujaya et al., 2013). This research is expected to provide a significant contribution to efforts to combat hate speech in Indonesia and become a reference for further research in this field.

RESEARCH METHOD

The research methodology involves the exploration of the Long Short Term Memory (LSTM) method, optimized using a Genetic Algorithm (GA), for detecting hate speech in Twitter data from Indonesia. The study employs Term Frequency-Inverse Document Frequency (TF-IDF) and Global Vectors for Word Representation (GloVe) as feature extraction techniques to create effective word vector representations in natural language processing. Additionally, feature expansion and the construction of a similarity corpus are introduced to enhance the performance of the LSTM classification model.

The evaluation of the model's performance is conducted using a confusion matrix, which measures key metrics such as accuracy, precision, recall, and F1 score. The research findings indicate that the LSTM model, when combined with TF-IDF and GloVe for feature extraction, achieves the highest accuracy, reaching up to 92.91%. Furthermore, the study identifies that the combination of Unigram, Bigram, and Trigram, with a maximum feature count of 10,000 and the GloVe corpus with the top 20 similarity scores, yields optimal results. The study also demonstrates that parameter optimization using genetic algorithms significantly improves both accuracy and F1 score.

This optimized LSTM model shows strong potential for accurately classifying hate speech in test data, which could be instrumental in the detection and management of hate speech on social media platforms. Moreover, it enhances the model's capability to identify and understand text content within the context of the Indonesian language.

RESULT AND DISCUSSION

Test Results

The study consists of four scenarios designed to systematically evaluate the LSTM classification model for the best performance, represented by accuracy values and F1Score which provides a better picture of the model's performance for unbalanced data. The accuracy value and F1Score are the average results of three executions to represent the overall accuracy. This is necessary because the Random State is not defined which causes performance variation with each model execution. The scenario can be seen in Table 1.

Table 1. Test Scenarios

Scenario	Description
1	Testing of the data ratio of the model separation and the best accuracy results are used as the next scenario.
2	Testing of models with *n-gram* combinations i.e. *unigram*, *bigram*, *trigram*, *unigram + bigram*, dan *unigram + bigram + trigram*.
3	Model testing by comparing data with 2,500, 5,000, 10,000 Feature vector.
4	Testing models with extension features using the similarity corpus that made from GloVe.
5	Model testing by optimizing parameters using *genetic algorithms*

Scenario 1

In the first scenario, the data sharing ratio was determined using tweet data on the unigram feature using the TF-IDF method as feature extraction with a max feature of 1000. The data sharing ratios are 70:30, 80:20, and 90:10. Meanwhile, the LSTM model used in this test is the best LSTM model obtained from the research of Helmi Imaduddin, Lucky Anggari Kusumaningtias and Fiddin Yusfida regarding the Application of LSTM and Glove Word Embedding for hate speech detection in Indonesia in 2023 (Imaduddin et al., 2023). The best model obtained consisted of LSTM layer with 32 units, dense layer with 6 units and the activation function used was ReLu, dropout layer with a dropout rate of 0.5, and dense layer with 1 unit and using the sigmoid activation function and the loss binary cross entropy function. The optimizer used is Adam optimizer with a learning rate of 0.001, a batch size of 128 and epochs of 10. The results of this test can be seen in Table 2.

Table 2. Scenario 1 Results

Ratio	Accuracy (%)	*F1-Score* (%)
70:30	88.68	88.99
80:20	88.75	88.94
90:10	**89.51**	**89.89**

The results of the first scenario were obtained that the 90:10 ratio resulted in the best accuracy performance and F1Score of 89.51% and 89.89%, respectively. Based on this scenario, split data with a ratio of 90:10 is used as the baseline for the next scenarios.

Scenario 2

The second scenario is performed to search for the best n-gram before performing feature extraction with TF IDF. The N-grams tested are Unigram (baseline), Bigram, Trigram, Unigram combination with Bigram, and Unigram, Bigram, Trigram combination. Table 3 shows the results of scenario 2.

Table 3. Scenario Result 2

N-Gram	Accuracy %	*F1-Score (%)*
Unigram (baseline)	89.51	89.89
Bigram	83.25 (-0.07)	82.21 (-0.085)
Trigram	67.23 (-0.249)	89.56 (-0.004)
Unigram+Bigram	89.53 (0)	89.56 (-0.004)
Unigram + Bigram + Trigram	89.64 (0.001)	89.90 (0)

The results of the second scenario show that Unigram + Bigram + Trigram results in an increase in accuracy and F1Score compared to other types of n-grams and has an increase in accuracy value against baseline of 0.1%. Based on this scenario, Unigram + Bigram + Trigram is used for the next scenarios using a baseline of 89.51%.

Scenario 3

The third scenario is looking for the best max feature from TF-IDF feature extraction. The max features tested include 2500, 5000 (baseline), and 10000. Table 4 is the result of scenario 3.

Table 4. Scenario 3 Results

Max Feature	Accuracy %	F1-Score (%)
5000 (baseline)	89.64	89.90
2500	88.41 (-0.014)	88.58 (-0.015)
10000	**90.42 (0.009)**	**90.70 (0.009)**

The results of the third scenario show that the max feature of 10000 results in an increase in accuracy and F1Score compared to other max features by 0.9%. Based on this scenario, the max feature of 10000 is used for the next scenarios.

Scenario 4

This scenario focuses on using Glove as a feature expansion. Testing was carried out using a corpus built using the Glove corpus. The top similarities used for feature expansion are Top 1, Top 5, Top 10, Top 15, Top 20, and Top 25. Table 5 is the test result of scenario 4.

Table 5. Scenario Results 4

1.1	**TWEET CORPUS**	**1.2**	**ACCURACY (%)**	***1.3***	***F1-SCORE***
BEST SCENARIO 3		1.4	90.42	1.5	90.70
1.6	TOP 1	1.7	89.8 (-0.007)	1.8	89.78 (-0.01)
1.9	TOP 5	1.10	90.17 (-0.003)	1.11	90.45 (-0.003)
1.12	TOP 10	1.13	90.34 (-0.001)	1.14	90.82 (0.001)
1.15	TOP 15	1.16	89.67 (-0.008)	1.17	89.65 (-0.012)
1.18	**TOP 20**	**1.19**	**90.66 (0.003)**	**1.20**	**91.10 (0.004)**
1.21	TOP 25	1.22	90.38 (0)	1.23	90.87 (0.002)

The results of the fourth scenario show that the corpus of tweets with the Top 20 similarity results in an increase in accuracy and F1Score compared to others which are 0.3% and 0.4%, respectively. Based on this, this model will use the best corpus for further testing, namely parameter optimization.

Scenario 5

Hyperparameters to be optimized

In this study, there are several hyperparameters that will be optimized, namely the number of LSTM neurons with values of 32, 64, 128, and 256, dropouts with values of 0.2, 0.35, 0.5, 0.65, and 0.8, the number of neurons in the dense hidden layer unit with values of 32, 64, 128, and 256, learning rate with values of 0.01, 0.001, and 0.0001, batch size with values of 16, 32, 64, and 128, and the number of epochs with values of 10, 15, 20, and 50 (Wei et al., 2021). Optimization of these hyperparameters is important to improve the performance of the hate speech detection model on Twitter and avoid overfitting or underfitting in the training process. The parameters to be optimized are shown in Table 6 as follows.

Table 6. List of parameters to be optimized

PARAMETER	VALUE
UNIT LSTM	32, 64, 128, 256
UNIT DENSE HIDDEN LAYER	32, 64, 128, 256
DROPOUT	0.2, 0.35, 0.5, 0.65, 0.8
LEARNING RATE	0.01, 0.001, 0.0001
BATCH SIZE	16, 32, 64, 128
EPOCH	10, 15, 20, 50

Genetic Algorithm (GA)

In this study, the fitness used is the accuracy value. After arranging chromosomes by generating 4 individuals consisting of combinatorics of each hyperparameter, then evaluate based on fitness function. Selection was carried out as many as 4 chromosomes from 4 parents from the population using the roullete wheel method. After obtaining the results of the initial population selection, crossover and mutation reproduction is carried out based on random numbers generated less than the probability value. The process is repeated from selection, crossover, to mutation until one of the stopping criteria is met. Choosing the best chromosomes by comparing fitness values in each generation. Table 7 shows the GA operating parameters used in this study.

Table 7. GA model parameter

PARAMETER	VALUE
CROSSOVER PROBABILITY	0.4
MUTATION PROBABILITY	0.1
SELECTION	SELROULETTE
POPULATION SIZE	4
NUMBER OF GENERATION	4
FITNESS FUNCTION	*ACCURACY*

The disadvantage of this LSTM-GA model is that the resulting solution can be different even though it uses the same GA parameter configuration for each training, so the resulting performance is also different each training is executed. This is because the genetic algorithm is a stochastic model that samples the population randomly, so the best parameters or solutions generated can be different because the population generated is different at each execution (Wiranata, 2021).

Genetic Algorithm (GA)

In this study, the fitness used is the accuracy value. After arranging chromosomes by generating 4 individuals consisting of combinatorics of each hyperparameter, then evaluate based on fitness function. Selection was carried out as many as 4 chromosomes from 4 parents from the population using the roullete wheel method. After obtaining the results of the initial population selection, crossover and mutation reproduction is carried out based on random numbers generated less than the probability value. The process is repeated from selection, crossover, to mutation until one of the stopping criteria is met. Choosing the best chromosomes by comparing fitness values in each generation. Table 8 shows the GA operating parameters used in this study.

Table 8. GA model parameter

PARAMETER	VALUE
CROSSOVER PROBABILITY	0.4
MUTATION PROBABILITY	0.1
SELECTION	SELROULETTE
POPULATION SIZE	4
NUMBER OF GENERATION	4
FITNESS FUNCTION	*ACCURACY*

The disadvantage of this LSTM-GA model is that the resulting solution can be different even though it uses the same GA parameter configuration for each training, so the resulting performance is also different each training is executed. This is because the genetic algorithm is a stochastic model that samples the population randomly, so the best parameters or solutions generated can be different because the population generated is different at each execution.

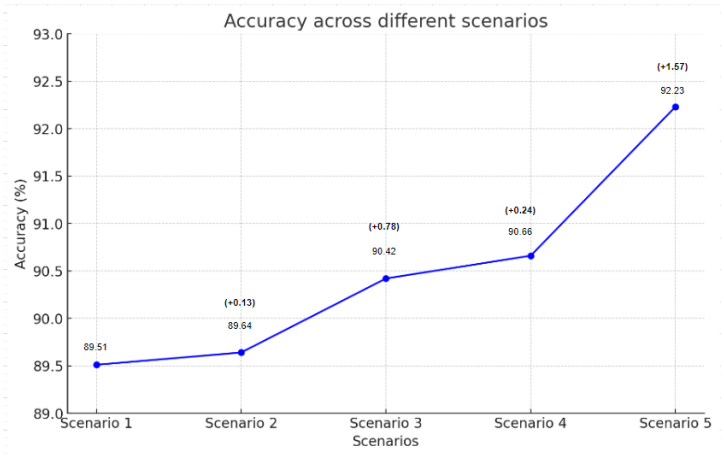

Figure 1. Increased Accuracy Score against baseline in all Scenarios

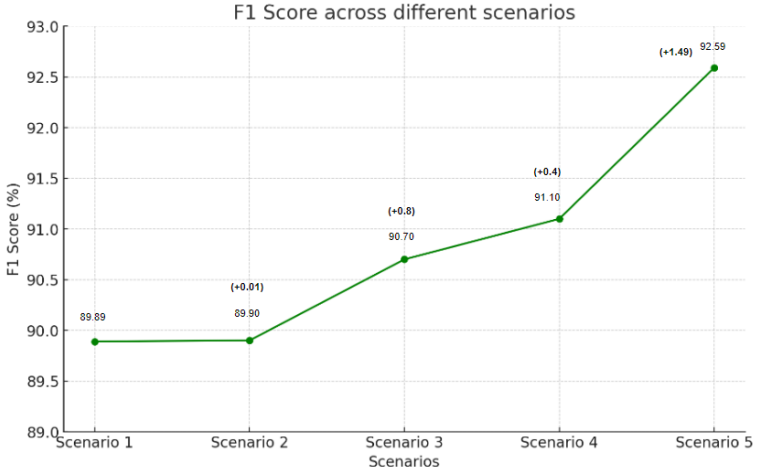

Figure 2. Increase of F1 Score to baseline in all Scenarios

Test Analysis Results

From the three models made using hyperparameters based on the best individual genetic algorithm results, the results of the classification of test data can be obtained which can be seen with the confusion matrix in table 9.

Tabel 9. Confusion Matrix Model LSTM

		Predicted	
		Buchan *Hate Speech*	*Hate Speech*
Current	Buchan *Hate Speech*	2145	249
	Hate Speech	93	2335

Based on Table 9, it can be seen that the number of hate speech tweets that are correctly classified is 2335 and hate speech tweets that are not properly classified are 93 tweets. Properly classified non-hate speech tweets were 2145 tweets and improperly classified non-hate speech tweets were 249 news. The following are the results of the accuracy calculation using the LSTM model.

$$\text{Accuracy} = \ = \ = 0{,}929075 \frac{TP+TN}{TP+TN+FP+FN} \frac{2335+2145}{2335+2145+249+93}$$

The results of the accuracy calculation show that the classification performance using the LSTM model is 92.91% which is able to accurately classify 4480 data out of 4822 total test data. The following is an analysis of prediction errors seen from the words contained in tweets in certain classes as follows.

Table 10. Best LSTM Model Predictions

Current	Prediction	Tweet Before Preprocessing	Tweet After Preprocessing
Hate Speech	*Hate Speech*	"@en_jeu_L @pengarang_sajak Loh loh loh kok ngamok, centang biru kelakuan berudu.ðŸ˜…	ngamok centang biru laku berudu orang orang era presiden jokowi puncak jaya orang orang goblok
Non Hate Speech	*Non Hate Speech*	Makan ketupat di depan cermin Pilihan 2024 Anies-Muhaimin #AMINPaslonSantri #AniesMuhaimin2024 @cakimiNOW @DPP_PKB https://t.co/zv15adSagI	makan ketupat cermin pilih anies muhaimin
Hate Speech	*Non Hate Speech*	Benar kata orang-orang ""Di era presiden Jokowi inilah puncak kejayaan orang-orang goblok"	duet anies sandi lengkap serta milu menang duet nyungsep cari pasang sandi
Non Hate Speech	*Hate Speech*	Makan ketupat di depan cermin Pilihan 2024 Anies-Muhaimin #AMINPaslonSantri #AniesMuhaimin2024 @cakimiNOW @DPP_PKB https://t.co/zv15adSagI	takut panggil polisi bego lempar besi besi empuk coba lempar besi kepala gila kali

Table 11. Highest Frequency of Words in Each Prediction Data Group

TN		TP		FN		FP	
Kata	Frekuensi	Kata	Frekuensi	Kata	Frekuensi	Kata	Frekuensi
anies	594	polisi	1136	anies	39	polisi	164
agama	519	kontol	860	polisi	25	bangsat	106
orang	246	bangsat	661	bangsat	22	orang	60
polisi	237	agama	591	agama	14	anies	60
tidak	172	kayak	486	kaum	14	agama	44
jokowi	143	orang	470	jelek	10	kontol	39
presiden	142	anjing	465	bacot	10	tidak	33
dukung	122	lgbt	393	orang	10	kayak	31
hebat	117	tolol	207	dukung	9	anjing	28
indonesia	116	muka	198	anak	8	islam	23

Based on the table above, it shows that tweets that are not hate speech that are correctly predicted as not hate speech (TN) mostly contain the words anies, religion, police, people, no, police, president, jokowi. Tweets that are correctly predicted as hate speech (TP) mostly contain harsh words. Meanwhile, tweets that are incorrectly predicted as not hate speech (FN) mostly contain words that are in the top words in true negative such as anies and police. Tweets that are not hate speech that are incorrectly predicted as hate speech (FP) mostly contain harsh words.

New Data Prediction

This stage is the stage of testing the model by giving a series of random texts to identify hate speech. This process involves predicting the model for each new text with the aim of classifying it as Hate Speech or Non-Hate Speech based on the context characteristics identified by the model. Table 12 is a tweet text that was taken randomly outside the trained dataset.

Table 12. New tweets outside the dataset

No	Tweet
1	ya gimana ya.. average of tiktok user itu bocah" yang fomo 😩 😩 ga heran lagi sih gua
2	Namanya demokrasi, terima aja hasil akhirnya. Kalau nanti mereka ngeluh ya senyum aja dulu baru ditolong Habis itu dorong ke jurang
3	Loh iya makanya dia memelihara ketololan tsb, biar gampang kampanye nya. Itu letak bangsatnya. Makanya cara mengentaskan ketololannya pun ngawur.

Based on Figure 3, this test shows three of them are identified as Non-Hate Speech, while the other three are identified as Hate Speech. These results show the model's ability to detect texts that contain or do not contain hate speech.

	Tweet	label
0	ya gimana ya.. average of tiktok user itu boc...	Non-Hate Speech
1	Namanya demokrasi, terima aja hasil akhirnya.\...	Non-Hate Speech
2	yg w gak terima tuh kita harus nurutin suara t...	Hate Speech
3	Loh iya makanya dia memelihara ketololan tsb, ...	Hate Speech
4	gabisa dibantah juga kalo user tiktok lebih ra...	Non-Hate Speech
5	By data emang tolol. Terima fakta. Justru kare...	Hate Speech

Figure 3 New Tweet Classification Results with LSTM Model.

The model is able to detect texts that contain or do not contain hate speech from the learning process (training) of the previous tweet dataset so that it is able to learn patterns for each characteristic. As previously explained, tweets that contain hate speech or do not contain hate speech have their own characteristics. For example, tweets that contain hate speech contain harsh words, so if there is a new text that contains harsh words, it will be predicted as a tweet containing hate speech.

CONCLUSION

In this study, we used the Long Short-Term Memory (LSTM) method to detect hate speech in Indonesian Twitter data. Researchers performed feature extraction using the TF-IDF and GloVe methods, and conducted evaluations using a combination of n-grams, Unigrams + Bigrams + Trigrams. In addition, researchers also optimized hyperparameters using genetic algorithms to improve the performance of the LSTM classification model. Glove uses 30 epochs, 100 components and a learning rate of 0.05 to create a corpus model. Based on research on hate speech that has been conducted on 48,920 Indonesian Twitter data using the LSTM model, five test scenarios were obtained. TF-IDF feature extraction is used at all stages of the test scenario. The expansion feature is also used to create a corpus with tweet data using the GloVe method. The first scenario proves that the best data separation ratio is obtained with a ratio of 90:10 with an accuracy of 89.51%, which is then used as the basis for the next scenario. The second scenario focuses on testing n-grams. The best performing N-gram is a combination of Unigram, Bigram, and Trigram with an accuracy of 89.64%. The third scenario shows that the best number of feature vectors is 10,000 features, so the maximum features in this test are used in the next scenario. The fourth scenario is to test feature expansion. The best accuracy is obtained on the tweet corpus with Top 20 similarity, which is 90.66%. The final scenario focuses on hyperparameter optimization using a genetic algorithm, the best individual produced is an individual with 256 LSTM units, 32 dense hidden layer units, a dropout rate of 0.5, a learning rate of 0.001, a batch size of 32 and an epoch of 10 with an accuracy of 92.23% with an increase of 2%.

REFERENCES

Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2018). Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, *11*(7), 1636.

Chui, K. T., Gupta, B. B., & Vasant, P. (2021). A genetic algorithm optimized RNN-LSTM model for remaining useful life prediction of turbofan engine. *Electronics*, *10*(3), 285.

Gautam, V. (2021). A Real Time Analysis of Offensive Texts to Prevent Cyberbullying. *Databases Theory and Applications: 32nd Australasian Database Conference, ADC 2021, Dunedin, New Zealand, January 29–February 5, 2021, Proceedings*, *12610*, 152.

Imaduddin, H., Kusumaningtias, L. A., & A'la, F. Y. (2023). Application of LSTM and GloVe Word Embedding for Hate Speech Detection in Indonesian Twitter Data. *Ingénierie Des Systèmes d'Information*, *28*(4).

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), e5909.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Sameer, M. (2022). *Hate Speech Detection in a mix of English and Hindi-English (Code-Mixed) Tweets*.

Talita, A. S., & Wiguna, A. (2019). Implementasi algoritma long short-term memory (LSTM) untuk mendeteksi ujaran kebencian (Hate Speech) pada kasus pilpres 2019. *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *19*(1), 37–44.

Tanujaya, W., Dewi, D. R. S., & Endah, D. (2013). Penerapan algoritma genetik untuk penyelesaian masalah vehicle routing di PT. MIF. *Widya Teknik*, *10*(1), 92–102.

Wei, B., Li, J., Gupta, A., Umair, H., Vovor, A., & Durzynski, N. (2021). Offensive language and hate speech detection with deep learning and transfer learning. *ArXiv Preprint ArXiv:2108.03305*.

Wiranata, R. B. (2021). A Genetic Algorithm Hyper-parameter Optimization of Ensemble Approach: Strategi Prediksi Saham Mempertimbangkan Indikator Teknikal & Sentimen Berita. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *8*(3), 1442–1456.