Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 1 (2025) 137 - 145 e-ISSN: 2580-0760

Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF

Syaikha Amirah Zikrina[1*], Fitriyani[2]
[1]Department of Data Science, Faculty of Informatics, Telkom University, Bandung, Indonesia
[2]Department of Informatics, Faculty of Informatics, Telkom University, Bandung, Indonesia
[1]syaikhaaz@student.telkomuniversity.ac.id, [2]fitriyani@telkomuniversity.ac.id

Abstract

Most of the hate speech and abusive content on social media, particularly in the Indonesian language, presents significant challenges for content moderation systems. Previous research has applied machine learning models such as Recurrent Neural Networks (RNN), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN) to address this issue. However, these approaches are limited in their ability to capture the relational and contextual nuances inherent in the data, resulting in suboptimal performance. This study introduces an approach by combining Graph Neural Networks (GNN) with Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction to improve hate speech detection on Twitter (platform X). The dataset consists of 13,169 Indonesian tweets, manually labeled for Hate Speech and Abusive categories. Preprocessing steps include text cleaning, stemming, stop-word removal, and normalization. The GNN model achieved superior results, with accuracy scores of 92.90% for Abusive and 89.78% for Hate Speech, significantly outperforming the RNN model, which achieved accuracy of 86.09% and 86.15%, respectively. This study highlights the advantage of graph-based approaches in capturing complex relationships within text data. Future research can explore expanding datasets to include regional dialects and integrating advanced feature extraction techniques like Word2Vec or BERT. This study establishes a robust framework for improving hate speech detection, offering a valuable contribution to safer digital environments.

Keywords: Context-Aware Sentiment Analysis; Graph Neural Network (GNN); Hate Speech Detection; Social Media; XTF-IDF

How to Cite: Syaikha and Fitriyani, "Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 1, pp. 137 - 145, Feb. 2025.
DOI: https://doi.org/10.29207/resti.v9i1.6179

1. Introduction

Many Indonesians are active users of Twitter, now known as platform X. The platform X has been extensively studied across various domain, techniques and topics [1]. With over 24 million users, platform X is the fifth most popular social media platform in Indonesia [2], [3]. It has become popular as a platform for sharing information, expressing opinions, and interacting with others. The large number of users and the freedom of speech make it more likely that abusive language in the form of hate speech will emerge [2]. Hate Speech Abusive Language (HSAL) refers to any form of communication that is intended to express hatred against a particular group based on their race, ethnicity, religion, gender, sexual orientation, or other identities. Hate speech can trigger social conflict, SARA, and even prompt violence [2], [4].

In recent years, certain internet users spread hate speech and abusive language (HSAL) on social media [5], [6]. Even though not all statements or expressions in abusive language contain offensive or dirty language. It implies that hate speech is not always synonymous with abusive language. Jokes and casual chats are frequent venues for using abusive language to convey affection [6]. Later, the government adopted the Electronic Information and Transaction Law (ITE), article 28

Received: 19-11-2024 | Accepted: 23-01-2025 | Published Online: 16-02-2025

paragraph 2 on hate speech, to handle or prevent these phenomena [7].

Sentiment analysis of hate speech on Platform X in the Indonesian language is particularly challenging due to the prevalence of coarse, informal, and slang expressions. These linguistic features are often difficult to identify and classify using existing tools developed by researchers [8]. This process involves inspecting an opinion to detect feelings, views, emotions, expressions, beliefs, attitudes, and opinion [9], [10]. For example, Imamah et al's [11], analyzed sentiments by collecting reviews of The Body Shop Tea Tree Oil on Female Daily Application. The study used pre-processing, and training with the Naïve Bayes algorithm, achieving an accuracy of 80,61%, Logistic Regression with an accuracy of 82,47% and an SVM accuracy of 83,71% [12]. Another study on sentiment analysis, by Kosasih et al [13]. It classified responses for online toy stores using K-Nearest Neighbor and TF-IDF, analyzing 1000 reviews of toy products and achieving an accuracy of around 79,33%.[13] . The TF-IDF method was employed to evaluate the importance of words within the documents, distinguishing between significant and less significant terms.

Another machine learning method is Graph Neural Network (GNN) which used to process data in a graph structure and to deliver satisfactory learning outcomes in graph representation [14],[15],[16]. GNN since its debut, has improved in effectiveness and efficiency that are now crucial for several applications, including developing recommendation systems and forecasting protein interactions [17]. As demonstrated by Nurfiqri et al [15], of GNN outperforms CNN and SVM in cyberbullying detection. GNN's achieving accuracy of 92,78% ability to model contextual and relational semantics within a graph structure. less than SVM [15], [18]. The compute time of GNN when compared to CNN, requires a significant 12 seconds time reduction.

Classification tools in previous studies in sentiment analysis found several disadvantages [19]. Such as numerous false negative errors, likely caused by the unbalanced dataset. An unbalanced dataset can give negative results on classification performance [7]. Meanwhile, Utami et al. [20] handled the unbalanced data with a combination of synthetic minority oversampling techniques (SMOTE) and Recurrent Neural Network (RNN) to analyze the sentiment of Shopee application user reviews.

According to DiPietro and Hager [21], RNNs under the category of deep learning are used to form a Neural Network for processing sequences, and they can employ distributed word representation by converting each token into matrix-forming vectors. They will store past data to figure out the data patterns [11], [12], [21]. The performance result was quite good achieving 80% of accuracy, 84.10% of precision and 88,10% of f1-score handled without preprocessing [20]. However, in 2018, Saksesi et al. [22] performed research on hate speech classification, which can classify the existence of Hate speech with an average precision of 91%, recall of 90% and accuracy of 91% by using the RNN algorithm [22].

Analyzing sentiment in the Indonesian language presents unique challenges due to the frequent use of coarse, informal, and slang expressions. However, some of the machine learning that has been proven by research has a great number of accuracies. While GNN and RNN give the best accuracy compared to others. Conversely, previous research using GNN and RNN produced more accurate results when compared to classification methods like KNN, SVM, Logistic Regression, and Naive Bayes. Even so, most of the predictions are derived from the nearest KNN. However, the results of the GNN and RNN classifications are not very good [23]. Studies that used SVM classification with an emphasis on hyperplane data analysis aim to improve semantic relationships in texts with noise and ambiguity. According to research by Joshi et al. [8] and Kosasih et al.[13], while SVM is rather effective in classifying texts, this model has a more complex ability to capture contextual relationships.

DiPietro and Hager's research [21] shows that RNNs can overcome flat sequences, but do not fully understand complex relationships between words, especially in informal or poorly structured language in contrast to GNNs which allow capture of a wider context and relationships between elements in complex texts. For example, research by Nurfiqri et al. [15] shows that GNNs are superior in detecting cyberbullying compared to CNNs and SVMs, achieving higher accuracy and being able to model relational context in graphed data. GNNs are compatible, words can be represented as vectors using methods like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec, GloVe) [24]. This indicates that GNNs are more effective in handling data that has relationships between elements that cannot be understood linearly by other methods such as RNNs and SVMs. His previous research by Zhang et al. [18] and Ardiani et al. [17] showed that GNNs proved to be effective in text analysis applications, such as short text classification and fake news detection.

However, when confronted with the informal vocabularies, slang, and occasionally context-specific semantics that are challenging to interpret in Indonesian tweets, the gap in the analysis of hate speech on Platform X remains. This study uses GNN with TF-IDF for feature extraction to close this gap. This approach maximizes the discovery of pertinent terms while simultaneously accounting for the intricate relationships that still exist between concepts. Compared to traditional techniques like RNN and SVM, this approach is more significant since it models contextual semantics in brief, noisy texts.

2. Research Methods

For this investigation, an Indonesian dataset with 13,169 rows and 13 columns was used, and formatted in a CSV file. The use of Python with Visual Studio Code and Google Collaboratory as the primary development environments was part of the analysis. Following dataset preparation, TF-IDF was used for feature weighting, and then Graph Neural Networks (GNN) were used for data classification. Figure 5 depicts the study process, which begins with data collection and continues with preprocessing to preserve the original text's integrity while preparing it for analysis. There are several primary steps in the preprocessing phase, which are covered in the sections that follow.

2.1 Dataset

The dataset used in this study was obtained from Kaggle. This is shown in Figure 1 giving the information of the dataset that includes two primary labels, Hate Speech (HS) that refers to tweets containing content that have discrimination or violence towards individual or groups based on their identity. Such as race, religion, ethnicity or gender.

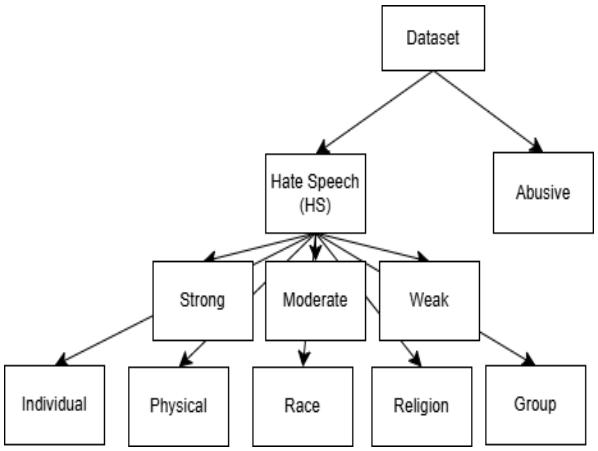

Figure 1. Dataset

Both tags were manually assigned by annotators using the binary Indonesian system, where 0 denotes the lack of a label and 1 denotes its presence. The annotators received explicit training on how to differentiate between characteristics like hate speech and non-hate speech to maintain consistency and accuracy when classifying [6].

Table 1. Dataset Label

Tweet	Hate Speech	Abusive
Bangkai apa om?	0	0
Benci sekali dengan umat islam	1	0
Kampang memang!!!	1	1
Blur kampret	0	1

The booklet contained common words, contextual cues, and a distinction between abusive language and informal discourse. As a result, the labelling procedure minimizes subjectivity and accurately reflects the tweet's purpose [25].Table 1 presents the label dataset used in this study. The dataset consists of stages of tweets labelled for two categories: Hate Speech and Abusive. Each tweet is said to be 0 and 1, where 1 indicates that there is hate speech or abusive content. while 0 indicates the opposite.

There is another example, bangkai apa om' which is not included in the category of Hate Speech and Abusive Relationships, which is written with a value category of 0 for both categories. Another example, 'Benci sekali dengan umat islam', includes the Hate Speech classification but not in the Abusive category which is labelled 1 for the Hate Speech category and 0 for the Abusive category. Another example is the tweet 'kampang memang' has been declared listed as 1 for both categories, which means that both have been strengthened in that category. And 'blur kampret' is labelled 0 for hate speech but listed for abusive.

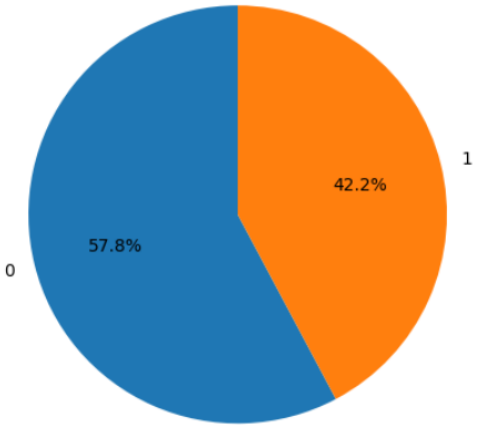

Figure 2 Hate speech (HS) Dataset Visualization

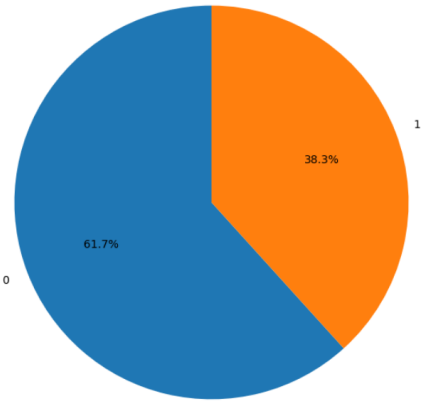

Figure 3 Abusive Dataset Visualization

Figure 2 and Figure 3 present the distribution of tweets for both categories, which include two pie charts comparing the frequency of data labelled "Abusive" and "HS". From the visualization, it is evident that the frequency of "Abusive" is lower than that of "HS." The terms "HS" and "Abusive" were chosen for analysis due to their interrelated nature. In Figure 2, the Abusive Dataset Visualization specifies that "Abusive" refers to content containing harassment or threats, while "HS" denotes content related to hatred or violence. Where

both of the visualization give 1 as a yes, and 0 means no including the Abusive nor Hate speech.

2.2 Pre-Processing

The most crucial step in doing an efficient analysis is cleaning the data. Prior to being entered into the model, the raw data used in this study underwent a number of cleaning and structuring procedures.

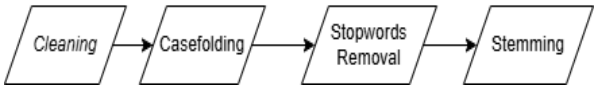

Figure 4 Preprocessing Steps

Figure 4, includes cleaning unnecessary symbols, and punctuation and eliminating irrelevant characters. The next step is case folding while in this step the text is normalized to lowercase, stop-word is the next step where the text removes common but uninformative terms, and the last is stemming, where in this step the text breaks down to their most basic form of word and handling informal language and colloquialisms, which are prevalent in Indonesian tweets.

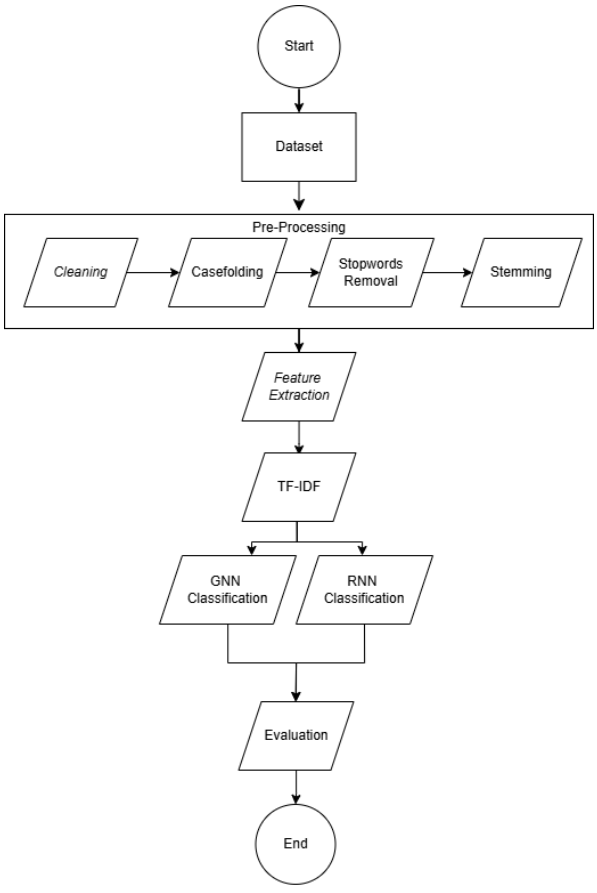

Figure 5 Research Flow

After preprocessing, the text data is converted into numerical features using the TF-IDF (Term Frequency-Inverse Document Frequency) method, which represents the importance of each word within the overall dataset. After that, these features are supplied into two machine learning models that classify the processed data, GNN (Graph Neural Network) Classification and RNN (Recurrent Neural Network) Classification. Finally, to evaluate the two models' performance and ascertain how well they identify hate speech and abusive content in tweets written in Indonesian, using metrics such as accuracy to determine the most effective model.

2.3 Feature Extraction TF-IDF

TF-IDF is a technique to evaluate the importance of a word in the text and its rarity in the corpus for feature extraction [26]. The TF-IDF score generated for each phrase in each tweet is integrated into the feature vector representing the tweet during the machine learning sentiment analysis process [26] Where there's research using KNN and TF-IDF methods with the NLP approach, using 260 reviews and gaining accuracy of around 77%, precision of 80% and recall of 74% [13].

$TF(t,d)$

$$\frac{Number\ of\ occurrences\ t\ in\ the\ document}{Total\ number\ of\ words\ in\ the\ document} \quad (1)$$

IDF(t,D)

$$log\frac{total\ number\ of\ documents}{1+number\ of\ documents\ containing\ the\ term\ t} \quad (2)$$

$$\text{TF-IDF (t,d,D)} = \text{TF(t,d)} \times \text{IDF(t,D)} \quad (3)$$

Feature extraction is a crucial step for classification, in this study the dataset was divided into training and testing sets using an 80:20 ratio to ensure the proper evaluation of the model on unseen data. The training set comprised 10,525 samples, while the testing set included 2,634. Those text samples were transformed into numerical representations using TF-IDF. TF-IDF process involves two main steps. First, it calculates the frequency of each word appearing in each document or text entity, referred to as Word Frequency (TF). Second, multiply the TF value by the Inverse Document Frequency (IDF) value to determine how important the word is in the document as a whole to Equations 1, 2 and 3.

TF-IDF results are used to calculate the result of word weighting which can be applied to the Graph Neural Network (GNN) model. In this experiment, TF-IDF was optimized using only the most pertinent phrases for the objective of understanding hate speech, with a feature count of 3000. In addition to lowering the computational overhead, this made sure that the qualities most likely to differentiate between hate speech and non-hate speech were kept to a minimum. Because the input used to generate the graph is cleaner and more targeted, this fine-tuning in the application of TF-IDF therefore enhanced the performance of the entire GNN model.

2.4 Graph Neural Network (GNN)

Graph Neural Network (GNN) refers to a type of model that can be applied to interpret data structures based on graph representations. GNN has become a useful tool for understanding the relationships between entities in

social networks and for classifying sentiment from the associated text [27]. GNN works particularly well with data that has few labels and undefined features [28], and it can reduce noise while highlighting significant characteristics by leveraging edges to connect related nodes.

$$h_{(i,j)} = f_{edge(h_i,h_j,x_{(i,j)})} \quad (4)$$

$$h_i' = f_{node(h_i,\Sigma_j \epsilon Ni)}, h_{(i,j),} Xi \quad (5)$$

Ni is the set of neighboring nodes that come to the i[th] node, and f_{edge} and f_{node} are two or three-layer Multilayer Perceptron (MLPs) that accept as input a set of parameter functions [29]. However, there are various other possible choices. In addition, multiple messaging updates can be chained by changing the $h_i \leftarrow h_i'$ after each node update which is determined by Equations 4 and 5. f_{edge} and f_{node} together are not required for message delivery updates [29]. The concept of algorithms known as Graph Neural Networks (GNNs) emerged in 2005. However, it is only in recent years that they have started to be used widely.

During the last few years, Graph Neural Networks (GNNs) have obtained outstanding performance across many deep learning tasks, with several variants such as Graph Convolutional Networks (GCNs), Graph Attention Network (GATs), and Graph SAGE [30]. GNN can exploit the relationships between words in the text, which are represented as graphs, it was chosen for this investigation. With this method, the model can capture considerably more intricate context patterns, like relational semantics and word-to-word syntax, which are frequently found in Indonesian hate speech data on Twitter.

This research adopts GCNs, a GNN variant to process the relationships between words in Indonesian hate speech tweets, represented as graphs. The main premise of the GCNs is used to encode the syntactic structure of sentences [28] the ability to extend the convolution operation in a graph domain. GCNs generally contain a fixed number of layers stacked against one another and in each layer, convolution and aggregation steps are performed for the improvement of the node embedding within the graph. Each of these layers allows the GCN to learn more and more complex structures and relations of the graph data [31], [32].

2.5. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a kind of neural network with memory states that can handle multiple inputs. A recurrent node receives input primarily from its own output. Since the data is processed automatically and without defining its features, RNN belongs to the category of deep learning[12]**.** The RNN calculation is shown in Equations 6 and 7.

$$S_t = \tanh(U_{xt} + W_{st-1}) \quad (6)$$

$$\check{y}_t = softmax\ V_{st} \quad (7)$$

In the analysis of feelings carried out by previous researchers using the RNN method [20] Where Recurrent Neural Network (RNN) is one method that can be used. RNN can recognize data patterns based on previous memories because RNN does not just discard information from the past. RNNs could process sequential data, which can be text data, time series data, voice data, and so on [20]. Three main parameters are used to start the RNN (Recurrent Neural Network) model in binary classification to detect hate speech and abusive content. The parameters consist of input_dim, hidden_dim, and output_dim. Input_dim indicates the number of TF-IDF extracted features, and hidden_dim is the hidden dimension with a value of 16 which is intended to store information from each time step. For binary classification, output_dim is set to 1.

3. Results and Discussions

Once the classification process was completed, the performance of the model was evaluated using relevant assessment metrics, including accuracy, precision, recall, and F1-score. The GNN model demonstrated a stronger ability to classify the data accurately based on the extracted features and the relationships between the data. To provide comparative insights, the performance of the RNN model, which is commonly referenced in previous studies as yielding strong results in similar tasks, was also analyzed.

3.1 Result

The preprocessing and feature extraction prepared the dataset for analysis. Text preprocessing pipeline includes text Cleaning, Case Folding, Stemming, and Normalization. The implications of this research in the context of hate speech detection on social media.

Figure 6 Wordcloud of Tweets

Figure 6 visualizes the dominant terms appearing in the dataset. “Jokowi”, ”Indonesia”, “Islam”, “Presiden”, and “agama” represent the highest frequency order of words in the tweets for understanding the patterns in the dataset to detect hate speech and abusive words.

To further comprehend the dataset’s characteristics, it has investigated the distribution of tweet lengths as illustrated in Figure 10. Most tweets in the dataset have a maximum length of up to 100 characters, which helps us understand the range of tweet lengths in the dataset. The distribution of tweet length also enables us to help

design appropriate preprocessing steps and design feature extraction for modelling.

Then, the dataset was trained by the GNN model with 100 epochs with a consistent decrease in loss value as shown in Figure 7. The experiment results present that the loss (prediction errors) decreases gradually in each iteration, which indicates that the model will be getting better at predicting. In the 100th epoch, the loss value was recorded at 0.1251 in each column. After training the GNN model for 100 epochs, it has to evaluate the model's performance using accuracy, precision, recall, and F1-score metrics that are given in Table 2 for better understanding.

The GNN model's performance was improved by using two hidden layers and a learning rate of 0.01. This learning rate of 0.01 has been shown to produce successful weight updates without overshooting because it strikes a compromise between the speed of convergence and the stability necessary during the training process. The Adam optimizer's performance improvement at this learning rate encouraged the model to gradually shrink the loss function over multiple epochs.

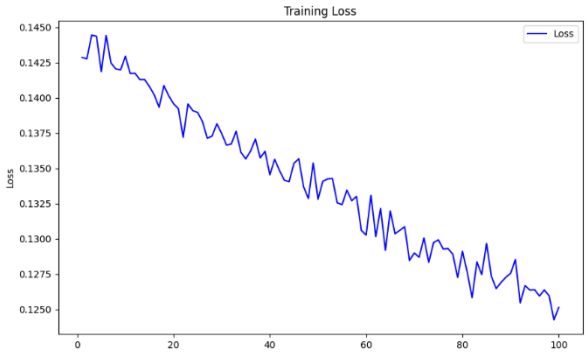

Figure 7 Training Loss

In the following research, another comparison method used is the Recurrent Neural Network model. RNN is chosen for comparison according to several references since this model gives the best results often for abusive content detection and hate speech. After conducting the test three times, it was observed that the performance of the RNN model was not too bad.

Table 2. Results of GNN

	Abusive	Hate speech
Accuracy	92,90%	89,78%
Precision	95,34%	94,43%
Recall	85,64%	80,56%
F1-Score	90,23%	86,94%

Table 2 shows the comparative performance of the GNN model, which performed better than the RNN model with high accuracy at the levels of hate speech (89.78%) and abusive speech (92.90%). Furthermore, GNN demonstrated its ability to manipulate relationally organized data by standing with the performance metrics of precision, recall, and F1-score. The durability of the GNN model was further demonstrated by variations in model parameters like learning rate and hidden layers; optimal performance versus two hidden layers was achieved at a learning rate of 0.01.

Table 3 shows the training process of the RNN model for the classification of abusive content and hate speech shows a consistent decrease in loss value from epoch to epoch. The loss value was recorded as 0.6648 in the first epoch, but it decreased gradually until it reached 0.0899 in the 100th epoch. This decrease in the loss value indicates that, with each iteration of the weight update, the model effectively reduces the prediction error. This indicates that the model is successfully learning the data patterns and relevant features for classification. While the RNN model seems to perform well, the data shown in Table 2 presents the case of how the GNN model outperforms the RNN model

Table 3. Results of RNN

	Abusive	HS (Hate speech)
Accuracy	86,09%	86,15%
Precision	82,36%	82,23%
Recall	80,06%	80,26%
F1-Score	81,20%	81,30%

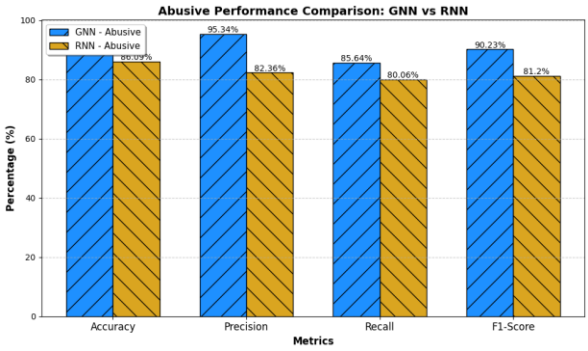

Figure 8 Abusive Visualization Performance

The results from the RNN model are promising enough and provide room for betterment in further works. The accuracy rate of 86.09% abusive and 86.15% hate speech can do well and catch most instances of this type. However, the precision and recall values retrieved, though commendable, also show areas that need to be upgraded for better performance. For instance, abusive content had an accuracy of 82.36%, and hate speech had 82.23%, which ipso facto means the residual probability of false positives in other words, misjudging non-abusive content continues to exist. In contrast, recall of 80.06% and 80.26% exhibited that some abusive content remains undetected, crucial for hate speech detection wherein overlooking those instances might have serious implications. The performance for abusive content detection, Figure 8. illustrates the performance. While the GNN model achieves higher accuracy at 92,90%, precision at 95,34%, recall at 85,64% and F1-Score at 90,23%. Outperforming the RNN model, which records an accuracy of 86,09%, precision of 82,36%, recall of 80,06%, and F1-score of 81,20%.

GNN has a more robust ability to accurately identify and differentiate abusive content. While Figure 7 shows the hate speech detection. Using the GNN model achieved another good performance with an accuracy of 89,78%, precision of 94,43%, recall of 80,56%, and an F1-score of 86,94%. Meanwhile, the RNN model gained an accuracy of 86,15%, a precision of 82,23%, a recall of 80,26%, and an F1-score of 81,30% as shown in Figure 9. This consistent performance advantage of GNN over RNN in both tasks suggests that GNN is more effective for hate speech and abusive content classification through the dataset.

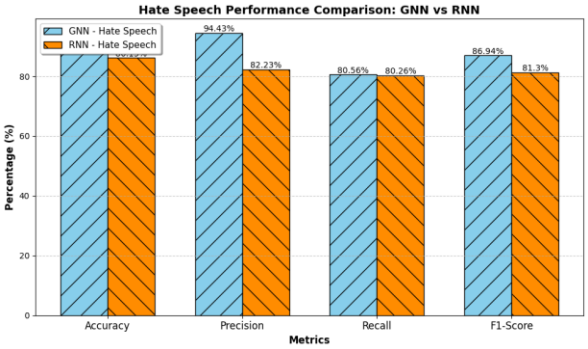

Figure 9. Hate speech Visualization Performance

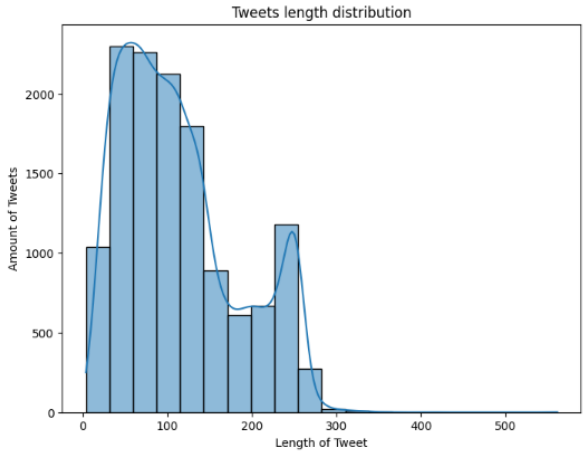

Figure 10. Tweets Length

Besides, from Figure 10, it is comprehensible that the length factor analysis of the tweets explains that most of the tweets are short, having a concentration below 100 characters, peaking from 50 to 100 characters. This nature signifies that platform X users express thoughts concisely, which is an important fact to know when preprocessing and feature engineering in sentiment analysis.

3.2 Discussions

The highly consistent performance of the GNN model over the RNN model shows that graph-based representations are effective for hate speech detection. With modelling relationships between various words and their contextual parameters, which usually go unnoticed by sequential models such as RNN to achieve a higher level of accuracy and F1-scores, the GNN is truly apt for the improvement of detection capability of hate speech. Also, due to the feature extraction of TF-IDF, significant terms really are strengthened, and a kind of improvement in robustness comes through the model as well.

Several key elements contributed to GNN's success in this investigation. A numerical representation of the relative importance of each word in the document, limited to 3000 characteristics, is first obtained by processing the text data using the TF-IDF approach. By doing this, the model is unable to learn from the irrelevant features. Second, cosine similarity is used to calculate the text similarity, which is then used to form the graph. Each node in this network represents a text (tweet), while edges are created between the texts based on a predetermined similarity criterion.

Additionally, by using a GNN architecture based on Graph Convolutional Networks, the model is able to capture intricate internal relationships of the graph, something that has proven challenging for earlier sequence-based techniques like RNNs. In order to greatly expand the inference of more contextual relationship patterns, the model has been built to incorporate and aggregate messages from surrounding nodes in a network. Training is made possible by this Binary Cross Entropy loss function, and the Adam optimizer aids in weight updates with learning rate stabilization and proficient learning. A set of criteria pertaining to accuracy, precision, recall, and F1 score in identifying distinct data categories has been used to guarantee balanced performance. When it comes to deciphering contextual and non-linear relationship patterns found in brief textual fragments, such as tweets, graphs can clearly outperform other data representation methods. This demonstrates that GNN can do better than other techniques, such as RNN, in simulating the intricacy of subtleties in hate speech data when preprocessing and using the best architecture design.

4. Conclusions

Adapting GNN architectures and TF-IDF for feature extraction, the research demonstrated that much of the hate speech can be detected in Indonesian tweets when combining the two sources. The preprocessing also refers to cleaning, stemming, stop-word removal, and normalization, which all lead to respectable and high-quality data for analysis. Additionally, the generation of a graph representation based on cosine similarity between TF-IDF vectors will enable the GNN model to capture relational and contextual information, making it perform better concerning the detection of the results depicted that the GNN model has recorded superior accuracy (92.90% for Abusive and 89.78% for Hate Speech) and F1-scores (90.23% for Abusive and 86.94% for Hate Speech) when both were compared to the RNN model. This result indicates the necessity of a graph-based approach in dealing with the challenges that are portrayed in detecting hate speech in more informal and multilingual datasets like Twitter. In addition, investigate potential areas for improvement.

One notable enhancement would be to make the model more generalizable, allowing it to include datasets containing tweets in various Indonesian dialects as well as those from other regions of Indonesia. Future investigations could look at alternative similarity measurements within different graph creation thresholds to see if performance improves. Furthermore, recent feature extraction approaches, such as word embeddings like Word2Vec or BERT, as well as advanced transformer models, will significantly improve the possibility of detecting hate speech and abusive content. Such effort will be required for future studies to improve and push the limits of the model's performance and applicability to various social media platforms.

References

[1] D. Murthy, "Sociology of Twitter/X: Trends, Challenges, and Future Research Directions," *Annu Rev Sociol*, vol. 50, no. 1, pp. 169–190, Aug. 2024, doi: 10.1146/annurev-soc-031021-035658.

[2] K. A. Rosyida and M. B. Siroj, "Strategi, Jenis Tindak Tutur dan Pola Tutur Pencemaran Nama Baik di Media Sosial," *Jurnal Sastra Indonesia*, vol. 10, no. 2, pp. 127–132, Jul. 2021, doi: 10.15294/jsi.v10i2.46672.

[3] I. Riadi, A. Fadlil, and U. Ahmad Dahlan Yogyakarta, "Identifying Hate Speech in Tweets with Sentiment Analysis on Indonesian Twitter Utilizing Support Vector Machine Algorithm," 2023.

[4] A. Matamoros-Fernández and J. Farkas, "Racism, Hate Speech, and Social Media: A Systematic Review and Critique," *Television & New Media*, vol. 22, no. 2, pp. 205–224, Feb. 2021, doi: 10.1177/1527476420982230.

[5] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, IEEE, Oct. 2017, pp. 233–238. doi: 10.1109/ICACSIS.2017.8355039.

[6] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 46–57. doi: 10.18653/v1/W19-3506.

[7] A. P. J. Dwitama, "DETEKSI UJARAN KEBENCIAN PADA TWITTER BAHASA INDONESIA MENGGUNAKAN MACHINE LEARNING: REVIU LITERATUR," *Jurnal Sains, Nalar, dan Aplikasi Teknologi Informasi*, vol. 1, no. 1, Aug. 2021, doi: 10.20885/snati.v1i1.5.

[8] S. Joshi, R. Dubey, A. Tiwari, and P. Jindal, "Sentiment Analysis Algorithms: Classifiers and Their Comparison," 2021, pp. 201–210. doi: 10.1007/978-981-16-1295-4_21.

[9] P. Liu, S. Joty, and H. Meng, "Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2015, pp. 1433–1443. doi: 10.18653/v1/D15-1168.

[10] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.10974.

[11] Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion," in *2020 6th Information Technology International Seminar (ITIS)*, IEEE, Oct. 2020, pp. 238–242. doi: 10.1109/ITIS50118.2020.9320958.

[12] Merinda Lestandy, Abdurrahim Abdurrahim, and Lailis Syafa'ah, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent Neural Network dan Naïve Bayes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 802–808, Aug. 2021, doi: 10.29207/resti.v5i4.3308.

[13] R. Kosasih and A. Alberto, "Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor," vol. 6, no. 1, 2021, doi: 10.30743/infotekjar.v6i1.3893.

[14] X. Ma *et al.*, "A Comprehensive Survey on Graph Anomaly Detection With Deep Learning," *IEEE Trans Knowl Data Eng*, vol. 35, no. 12, pp. 12012–12038, Dec. 2023, doi: 10.1109/TKDE.2021.3118815.

[15] Muhammad Rizki Nurfiqri and Fitriyani, "The Performance Analysis of Graph Neural Network (GNN) and Convolutional Neural Network (CNN) Algorithms for Cyberbullying Detection in Twitter Comments," *Indonesian Journal of Computer Science*, vol. 13, no. 3, Jun. 2024, doi: 10.33022/ijcs.v13i3.3940.

[16] "Penerapan graph neural network dalam pembangungan sistem rekomendasi."

[17] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *Jurnal Sistem dan Teknologi Informasi (Justin)*, vol. 8, no. 2, p. 183, Apr. 2020, doi: 10.26418/justin.v8i2.36776.

[18] B. Zhang, Q. He, and D. Zhang, "Heterogeneous Graph Neural Network for Short Text Classification," *Applied Sciences*, vol. 12, no. 17, p. 8711, Aug. 2022, doi: 10.3390/app12178711.

[19] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decision Analytics Journal*, vol. 3, p. 100073, Jun. 2022, doi: 10.1016/j.dajour.2022.100073.

[20] H. Utami, "Analisis Sentimen dari Aplikasi Shopee Indonesia Menggunakan Metode Recurrent Neural Network," *Indonesian Journal of Applied Statistics*, vol. 5, no. 1, p. 31, May 2022, doi: 10.13057/ijas.v5i1.56825.

[21] R. DiPitero and D. Hager, "Medical Image Computing and Computer Assisted Internation," 2014, pp. 503–514.

[22] A. S. Saksesi, "HS RNN," 2018.

[23] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 39–43, Jul. 2020, doi: 10.33096/ijodas.v1i2.13.

[24] K. P. Santoso, R. Venantius, H. Ginardi, R. A. Sastrowardoyo, and F. A. Madany, "Leveraging Spatial and Semantic Feature Extraction for Skin Cancer Diagnosis with Capsule Networks and Graph Neural Networks."

[25] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive Language and Hate Speech Detection for Indonesian-Local Language in Social Media Text," 2021, pp. 88–98. doi: 10.1007/978-3-030-79757-7_9.

[26] wesam ahmed, N. Semary, K. Amin, and M. Adel Hammad, "Sentiment Analysis on Twitter Using Machine Learning Techniques and TF-IDF Feature Extraction: A Comparative Study," *IJCI. International Journal of Computers and Information*, vol. 10, no. 3, pp. 52–57, Nov. 2023, doi: 10.21608/ijci.2023.236052.1128.

[27] H. Phan and A. Jannesari, "Story point level classification by text level graph neural network," in *Proceedings of the 1st International Workshop on Natural Language-based Software Engineering*, New York, NY, USA: ACM, May 2022, pp. 75–78. doi: 10.1145/3528588.3528654.

[28] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification," Sep. 2018, [Online]. Available: http://arxiv.org/abs/1809.05679

[29] L. Waikhom and R. Patgiri, "Graph Neural Networks: Methods, Applications, and Opportunities," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.10733

[30] F. B. Mahmud, M. Md. S. Rayhan, M. H. Shuvo, I. Sadia, and Md. K. Morol, "A comparative analysis of Graph Neural Networks and commonly used machine learning algorithms on fake news detection," in *2022 7th International Conference on Data Science and Machine Learning*

Applications (CDMA), IEEE, Mar. 2022, pp. 97–102. doi: 10.1109/CDMA54072.2022.00021.

[31] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," May 2016, [Online]. Available: http://arxiv.org/abs/1606.00061

[32] W. Fan *et al.*, "Graph neural networks for social recommendation," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, Association for Computing Machinery, Inc, May 2019, pp. 417–426. doi: 10.1145/3308558.3313488.

