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

Revue d'Intelligence Artificielle
Vol. 37, No. 6, December, 2023, pp. 1471-1481
Journal homepage: http://iieta.org/journals/ria

Short Text Classification Based on Hybrid Semantic Expansion and Bidirectional GRU (BiGRU) Based Method to Improve Hate Speech Detection

Ari Muzakir[1,2*], Kusworo Adi[3], Retno Kusumaningrum[4]

[1] Doctoral Program of Information System, School of Postgraduated Studies, Diponegoro University, Semarang 5024, Indonesia
[2] Faculty of Science and Technology, Bina Darma University, Palembang 30264, Indonesia
[3] Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia
[4] Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia

Corresponding Author Email: arimuzakir@binadarma.ac.id

https://doi.org/10.18280/ria.370611

Received: 25 September 2023
Revised: 18 October 2023
Accepted: 24 November 2023
Available online: 27 December 2023

Keywords:
short text classification, text expansion techniques, word sense disambiguation, semantic similarity measures, hybrid semantic expansion, hate speech detection

ABSTRACT

The persistent prevalence of hate speech on contemporary social media platforms demands advanced detection methods to address specific categories and levels of offenses. This research focuses on enhancing hate speech detection by refining text representation through a semantic expansion approach, surpassing the limitations of conventional methods. The back-translation technique is employed to enhance sentence structure. Initially, the Lesk Algorithm is utilized for word disambiguation in the semantic expansion process, identifying word meanings within relevant contexts. Subsequently, knowledge bases from WordNet and Kateglo are leveraged to enrich contextual information. The final step involves using Cosine Similarity to select the most appropriate words based on the highest scores. The combined semantic expansion technique significantly improves classification performance compared to conventional methods. Data, with and without semantic expansion, is vectorized into the BERT embedding space and classified using deep learning models such as CNN, BiGRU, and BiLSTM. The proposed approach consistently demonstrates high accuracy across all model types: CNN (88%), BiGRU (88.3%), and BiLSTM (87.3%). In contrast, models without semantic expansion yield relatively lower results-CNN (83.6%), BiGRU (83.3%), and BiLSTM (83.1%). This underscores the substantial breakthrough of the semantic expansion approach in overcoming challenges related to data distribution and semantic feature scarcity, ultimately resulting in improved classification performance.

1. INTRODUCTION

Hate speech on social media has emerged as a pervasive concern, largely attributable to the misuse of technology [1]. While interpretations of hate speech may vary, it is crucial to underscore that international law and religious doctrines unequivocally prohibit actions that incite hatred based on nationality, race, or religion, ultimately fostering hostility and violence [2]. The manifestation of hate speech on social media predominantly takes the form of concise textual expressions, such as comments, quotes, and posts, exhibiting an exponential growth trajectory. However, the inherent limitations of short texts, encompassing constrained features, contextual information deficits, ambiguity, and spelling errors, pose significant challenges to achieving satisfactory classification performance [3].

The task of text classification, particularly in the realm of short texts, is inherently complex and necessitates the discerning selection of appropriate methods and models. Short texts, with their inherent intricacies, often prove more formidable compared to their longer counterparts, thereby potentially leading to a degradation in classification quality, the introduction of bias, and suboptimal performance [4]. The accurate classification of short texts, ranging from quotes and social media tweets to comments and brief messages, assumes paramount importance in deciphering user intent [5]. Nevertheless, the classification of short texts frequently grapples with the absence of adequate context, giving rise to challenges such as data sparsity and ambiguity [6]. Conventional methodologies, exemplified by the bag-of-words (BoW) approach, often fall short of effectiveness in the realm of short texts [5], as they overlook word order and semantic relationships between words. Similarly, the Vector Space Model (VSM) proves less efficacious in the context of short texts, neglecting contextual nuances and the inherent significance of individual words, thereby resulting in information loss [3]. Consequently, as a viable resolution, researchers routinely pivot towards word embedding models such as Word2vec [7] and BERT [8] to address the intricacies associated with text representation.

Numerous studies have endeavored to confront these challenges through a variety of approaches. Chen et al. [9]

employed latent semantics and Latent Dirichlet Allocation (LDA) via Wikipedia for short text development. In a parallel vein, Wang et al. [5] fortified text representation by conducting comprehensive web crawls and integrating the acquired results into the realm of short text classification. Jelodar et al. [10], on the other hand, introduced the Biterm Topic Model (BTM) as an innovative LDA variant specifically designed to address the persistent issue of data sparsity in modeling short texts. Despite concerted efforts to refine our understanding of sentence context in the domain of short texts, the precise extraction of relevant information for targeted searches remains an ever-evolving research focus.

One strategy that has been deployed is text expansion, entailing the incorporation of word variations from a corpus with the aid of external knowledge sources. Although text expansion often introduces a degree of ambiguity due to the array of word choices available, the Lesk algorithm has demonstrated remarkable efficacy in mitigating sentence-level ambiguity. Naskar and Bandyopadhyay [11] convincingly demonstrated the prowess of the Lesk algorithm in addressing word ambiguity during information extraction from knowledge bases. Another noteworthy approach is Word Sense Disambiguation (WSD), which seeks to determine the precise meaning of target words within a given context. Chaplot and Salakhutdinov [12] introduced a knowledge-based WSD algorithm aimed at unraveling the meaning of each word across an entire document. However, the potential for context bias arises from the multitude of candidate words in text expansion. Hence, an efficient strategy involves the calculation of semantic similarity based on embeddings [13].

In their comprehensive research [14], the researchers leveraged WordNet to expand queries by identifying synonyms, extracting natural language phrases, and systematically sorting search results. Additionally, Azad et al. [15] proposed an innovative term weighting scheme for expansion terms, drawing insights from the extensive knowledge bases of Wikipedia and WordNet. Concurrently, Muzakir et al. [16] meticulously crafted a framework for hate speech detection, relying on robust knowledge bases such as WordNet and Kateglo to effectively alleviate ambiguity in short texts.

Numerous prior studies have explored various techniques and approaches for text expansion. For instance, Wang et al. [17] proposed the use of semantic units, extracting semantically closest word vectors to form an expanded matrix. This matrix is then projected into a CNN to derive global feature representations. Expanding on this work, Wang et al. [5] delved into a semantic group and word embedding approach, underscoring the idea that words with similar contexts exhibit closer vector representations. Additive semantic composition methods facilitated the computation of multiscale semantic units. In a different vein, Adhi et al. [18] achieved heightened accuracy using Naïve Bayes and semantic expansion, leveraging Kateglo knowledge for Twitter sentiment analysis. Additionally, Wang et al. [17], harnessed semantic expansion, utilizing the external knowledge base Probase to augment the meaning of short text. In a unique approach, Jahan et al. [19] introduced semantic expansion data, wherein synonymous words are extracted based on synonym groups from WordNet. This extraction process is coupled with an identification and disambiguation procedure to ensure the selected synonyms align with the sentence context.

Recent attention in Natural Language Processing (NLP) has shifted towards deep learning-based approaches. Specifically, models rooted in Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) have gained prominence for short text classification tasks [4]. Widely adopted architectures, such as GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory), retain memory from previous activations without fully overwriting values during processing, addressing limitations present in traditional RNNs [20]. In the context of this study, the BiGRU (Bidirectional GRU) architecture is employed to surmount challenges associated with RNNs, mitigate concerns about overfitting, and concurrently expedite training times [21].

This research aims to address challenges inherent in short text classification, where the frequent absence of context often leads to issues like data sparsity and ambiguity. The primary objective of this study is to explore semantic expansion techniques, specifically targeting the enrichment of contextual information within sentences. This exploration seeks to overcome limitations associated with conventional expansion methods. By implementing semantic expansion methods that utilize WordNet and Kateglo, the research aims to enhance classification performance through processes such as disambiguation and word expansion. The integration of Cosine Similarity in semantic similarity facilitates the precise measurement and comparison of terms for use in semantic expansion. Consequently, this research holds the potential to significantly contribute to the improvement of hate speech detection in short texts.

As such, this research is formulated around two primary inquiries: first, the application of semantic expansion techniques in short text classification, and second, the effectiveness of these techniques in detecting hate speech within the predominantly short-text-dominated landscape of social media.

To achieve these objectives, the research introduces a novel model employing a hybrid semantic expansion approach. This model is refined by addressing potential spelling errors in short texts and enhancing the coverage of the pretraining word vector table through back-translation methods. The hybrid approach plays a pivotal role in augmenting semantic understanding, capturing latent contextual meanings, and ultimately determining hate speech categorization within short texts. Furthermore, the research conducts a comprehensive comparison between data with semantic expansion and data without semantic expansion, utilizing classification methods such as CNN, BiGRU, and BiLSTM.

2. PROPOSED METHODOLOGY

Throughout this research, we propose implementing back-translation techniques as a strategic approach to enhance sentence quality. The process begins by generating synthetic data through back-translation, with the goal of refining language structure and rectifying any writing or spelling errors present in the original queries. Following this, we delve into the semantic expansion phase, leveraging knowledge from both WordNet and Kateglo to access a more extensive array of synonyms. The Lesk algorithm is then employed for word disambiguation, aiding in the identification of target words from the preceding process. Subsequently, semantic similarity scores are computed using the Cosine Similarity method. Finally, the model is trained using the BiGRU architecture to effectively classify hate speech content (Figure 1).

BiGRU's unique capability lies in enabling the model to comprehend the contextual nuances of a word or phrase by assimilating information from both antecedent (prior context) and subsequent (following context) directions. This dual-directional approach significantly aids the model in more accurately discerning the relationships and meanings of words within a specific context. Consequently, the BiGRU model proves instrumental in overcoming semantic challenges inherent in language processing.

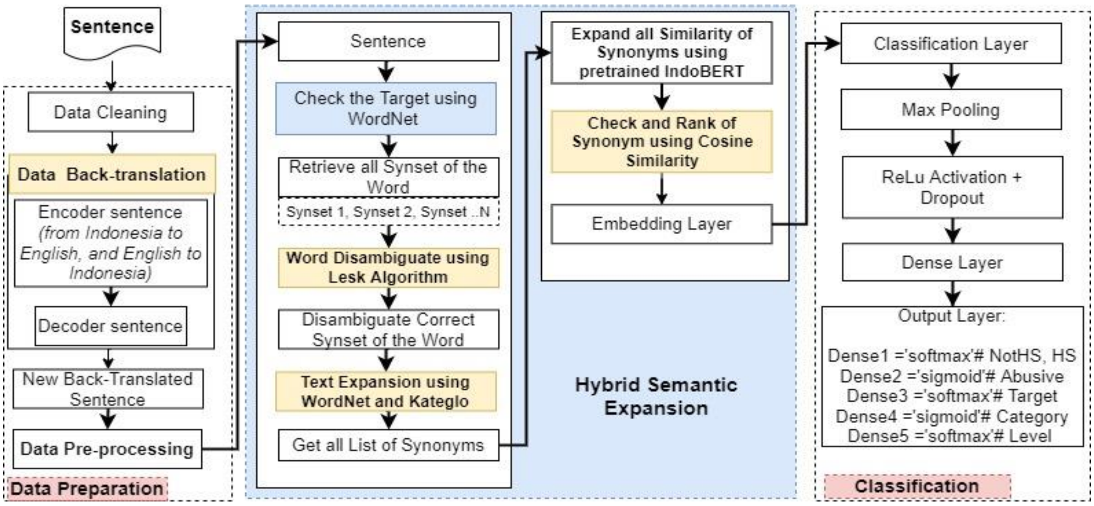


Figure 1. Proposed semantic expansion workflow for hate speech classification

2.1 Data

In the context of this study, we employed a dataset sourced from the GitHub repository [22]. Comprising 13,169 tweets gathered from Twitter, the dataset underwent meticulous annotation by 30 annotators. Characterized by a multi-label and multi-output structure, each sentence containing hate speech is systematically classified into three categories: target, category, and level.

Despite undergoing meticulous validation, the dataset reveals an imbalance in data distribution, particularly in the 'physical' and 'gender' labels, where positive tweets constitute less than 5 percent of the total dataset, as illustrated in Table 1. The primary challenge at hand is the observed data imbalance, exacerbated by the interdependence among label columns such as *'Hate Speech', 'Target', 'Category', and 'Level'*. While specific techniques to address data imbalance were not employed in this study, the clear necessity for careful data balancing is apparent.

Table 1. Data distribution for each label

Label	Positive	Negative
Not hate speech	7607	5561
Hate speech	5561	7607
Abusive	5043	8125
HS_individual	3575	9593
HS_group	1986	11182
HS_religion	793	12375
HS_race	566	12602
HS_physical	323	12845
HS_gender	306	12862
HS_other	3740	9428
HS_weak	3383	9785
HS_moderate	1705	11463
HS_strong	473	12695

2.2 Data back-translation and pre-processing

The Back-translation technique involves a two-step process: translating the text into another language and then rendering it back into the original language. This method generates new textual data with distinct vocabulary, yet it retains the original context and meaning. We applied this technique, as detailed in [24], to craft new sentences that encapsulate the core essence of the original sentences with enhanced grammatical structures. Employing back-translation not only rectifies language structure but also addresses writing and spelling errors, thereby augmenting the coverage of the pretraining word vector table. However, it is crucial to note that this technique may be unsuitable for certain sentence types. For instance, sentences containing slang words absent from the language corpus may result in interpretations deviating from the intended meaning.

Take the sentence "*Bro, it's really fun, I'm experiencing a lot of hype!*" In this scenario, the initial translation might fail to capture the subtleties of slang words like "*really fun*" and "*hype*" accurately. Upon back-translation, these slang words are likely to be interpreted literally, neglecting their genuine meaning or nuances within a specific social or cultural context. This limitation arises due to potential inadequacies in the translation model's understanding of slang vocabulary and the corresponding cultural context.

Moreover, during the preprocessing phase, every word in the dataset undergoes normalization, tokenization, and stop-word removal. This normalization is facilitated by a dictionary sourced from the same repository, which substitutes slang words with standardized forms predetermined [23].

2.3 Hybrid semantic expansion

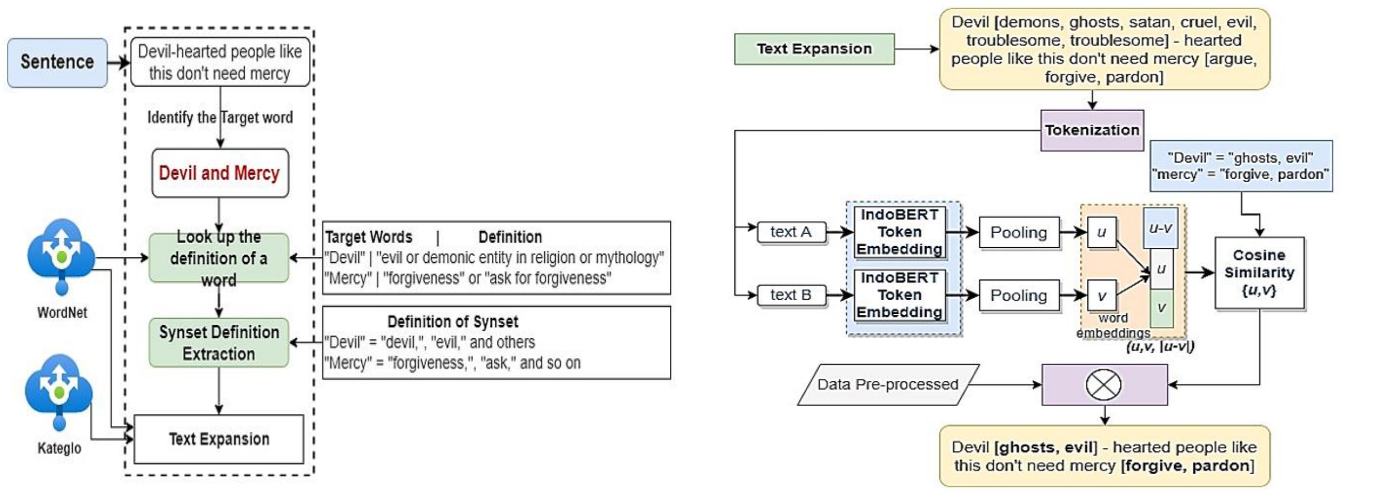
Short text expansion is a process with the primary goal of acquiring a broader context or more comprehensive information from texts composed of short phrases. This is achieved by leveraging external knowledge bases. This endeavor aims to address the ambiguity or information gaps frequently present in these concise expressions. In this context, the significance of short text expansion becomes paramount for identifying and elucidating the precise meanings of these phrases, as exemplified in Table 2.

Table 2. Examples of text expansion using wordnet and kateglo knowledge bases

Sentence	Text Expansion
Research institutions or provocative institutions... It's better to just disband them.	Institution ['organization', 'formation', 'association', 'corporation', 'realization'] research ['investigation', 'question', 'research'] or provocative agency. It's better to just break it up ['break up', 'disperse', 'scatter', 'end']!
People with devil hearts like this don't need to be forgiven.	Human ['individual', 'insan', 'human', 'person, 'someone'] devil-hearted ['evil', 'cruel', 'troublesome', 'troublesome', 'satan'] like this do not need ['need', 'require', 'request', 'desire', 'ask', 'require'] mercy.

Conversely, Word Sense Disambiguation (WSD) involves the intricate process of deciphering the actual meaning of words endowed with multiple interpretations or potential synonyms. By incorporating short text expansion and deploying word disambiguation techniques, we elevate the comprehension and processing of text, thereby mitigating misunderstandings and yielding more precise outcomes.

Illustrated in Figure 2 is a visual example of sentences undergoing the text expansion process with WSD and subsequent semantic similarity ranking. In instances involving target words, we perform calculations to compile a list of pertinent synonyms, drawing from resources such as WordNet and Kateglo. This meticulous step aims to ensure the selection of only the most contextually fitting synonyms. To achieve this, we input the synonym set and the original sentence containing the target word into the Lesk Algorithm, effectively eliminating ambiguous words. Following the disambiguation stage, we generate new sentences by incorporating the target word with each selected synonym, resulting in broader and more detailed sentence structures (Figure 2a). From these expanded sentences, we leverage pre-trained IndoBERT embeddings to guarantee semantic similarity and quantify the degree of likeness between two vectors using Cosine Similarity (Figure 2b).



(a) (b)

Figure 2. Ilustration of hybrid semantic expansion: (a) text expansion through word disambiguation and (b) semantic similarity using cosine similarity

Renowned for its efficiency in grasping word semantics, word embedding has found widespread application in Natural Language Processing (NLP), particularly for word sense disambiguation [24]. Within the embedding space, neighboring words exhibit semantic relationships [25], although the precise extent of these semantics remains unknown beforehand. Encompassing a broad vocabulary, word embeddings also excel in capturing a variety of syntactic and semantic connections [26].

In this study, an evaluation of semantic similarity was conducted for all text expansions using a pre-trained BERT model, further refined through fine-tuning with the IndoBERT, IndoBERT-Tweet [27], and BERT-Multilingual models. The primary objective of assessing semantic similarity is to gauge the degree of resemblance between words based on their conceptual semantic relationships, employing semantic similarity metrics [28, 29]. A common method for measuring semantic similarity is cosine similarity [13], which assesses the likeness between two sentences by tallying the shared words within them. Eq. (1) explicates the use of word vectors in calculating cosine similarity.

$$\mathrm{Cos}(S1,S2)=\frac{S1.S2}{||S1||.||S2||}=\frac{\sum_{i=1}^{k} S1_i S2_i}{\sqrt{\sum_{i=1}^{k} S1_i{}^2 \sum_{i=1}^{k} S2_i{}^2}} \tag{1}$$

In the context of vector representation for candidate sentences (S) and source sentences (S1) in dimension k, a comprehensive comparison is made between each sentence in two target documents and all sentences in the source document. The primary goal is to identify semantic relationships between target candidates and source sentences, utilizing the Cosine Similarity method. This method quantifies the degree of word similarity between the source sentence (S1) and the target candidate sentence (S2), with similarity scores computed based on word vectors [13].

2.4 Fine-tunning BERT

BERT has showcased remarkable performance in Natural Language Understanding (NLU) tasks, excelling in its ability to operate with less data, hastening development, and delivering superior results [8]. The IndoBERT models, specifically tailored for the Indonesian language by Wilie et al.

[30], play a pivotal role within the IndoNLU framework, catering to training, evaluation, and comparison needs, particularly in text classification tasks. IndoBERT stands out as the most sophisticated BERT-based language model for Indonesian, with pre-training objectives encompassing masked language modeling (MLM) and next sentence prediction (NSP).

In this experimental context, we leverage three pre-trained models: IndoBERT$_{\text{BASE}}$ *(indobenchmark/indobert-base-p2)*, IndoBERT-Tweet *(indolem/indobertweet-base-uncased),* and BERT-Multilingual$_{\text{BASE}}$ *(bert-base-multilingual-uncased)*. Each model undergoes training on TPUs, with a maximum sequence length of 256. BERT classification training involves input ids, attention masks, and labels from the merged training dataset, structured into a TensorDataset. Subsequently, samples in the *TensorDataset* are randomly divided into training and validation sets, maintaining an 80:20 ratio. This division is carried out with the aim of maximizing available data samples for both training and validation. The IndoBERTBASE model comprises 12 transformer layers, 12 attention heads, an embedding size of 768, and a hidden size of 768.

In this study, we enhanced IndoBERT for the classification task using Tesla A100 and V100 GPUs, equipped with 16GB GPU RAM and 51GB system RAM, accessible on Google Colab Pro. We selected the *BertForSequenceClassification* model, configured with a single layer for the classification task. During the training and validation phase, sequences are constrained to 80 tokens, shaping the development of the BERT classifier. For the pre-trained BERT phase, fine-tuning involves adjusting hyperparameters such as optimizer=Adam, batch_size=16, epoch numbers of 20 and 30, and learning rates (LR) with values of 0.001, 0.0001, 2e-05, 3e-05, and 5e-05. While LR values of 2e-05, 3e-05, and 5e-05 have been previously recommended, suboptimal results prompted the addition of LR values 0.001 and 0.0001 [31], ultimately achieving the best results with LR 0.001. Additionally, we opted for a batch size of 16, following guidance from other sources [31] that suggests larger batch sizes can expedite computations but may yield suboptimal training outcomes. Meanwhile, Devlin et al. [8] recommends a batch size of 16 and 32.

2.5 BiGRU classifier

Deep learning presents advantages over traditional machine learning, as it often requires less human intervention for feature extraction [32]. While traditional approaches like Naive Bayes (NB), Support Vector Machine (SVM), NBSVM, Latent Dirichlet Allocation (LDA), and Artificial Neural Network (ANN) struggle with short texts containing special characters and are susceptible to data sparsity and a lack of semantic features [33], deep learning methods, such as Convolutional Neural Networks (CNN), face challenges in effectively capturing contextual information. Models like Long Short-Term Memory (LSTM), despite their ability to consider context, may experience slow convergence. As an alternative, the Bidirectional Gated Recurrent Unit (BiGRU) model, featuring a bidirectional sequential structure, has proven effective in overcoming these challenges. The reset gate mechanisms, representing the level of neglect of the output of hidden layer neurons at the previous time step, enable BiGRU to reduce information loss due to increased reset gate speed. It is important to note that the direct application of the BiGRU model can result in high computational loads due to its involvement with high-dimensional input [34].

In our classification task, we adopted the BiGRU architecture based on the research by Qiu et al. [34], offering flexibility with one or multiple layers. Through experimental comparisons, we determined the optimal number of layers to understand its impact on the model. The experiments involved multiple layers of bidirectional GRU with 64 parallel operating neuron units. Subsequently, the BiGRU model underwent pooling through GlobalMaxPooling layers and was connected to a fully connected layer (Figure 3). Finally, the Dense output layer was configured with five outputs for hate speech classification (Table 3). Using the cross-entropy loss function to evaluate the model's performance, we successfully achieved accurate classification results. Additionally, we conducted a comprehensive comparison between the BiGRU, CNN, and LSTM architectures, employing various embedding models and datasets, both with and without semantic expansion. For the CNN and LSTM architecture settings, we followed the methodology presented in the study by Ependi et al. [35].

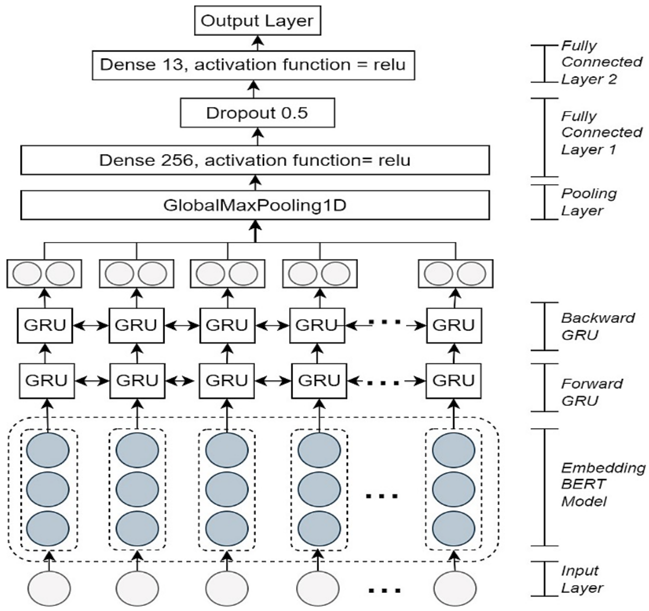


Figure 3. BiGRU model architecture

Table 3. Dense layer configuration

Output Layer	Description
Dense(1, ‘sigmoid')	For the labels of not hate speech (HS) and HS
Dense(1, 'sigmoid')	For the label of abusive
Dense(2, 'softmax')	For the target HS (individual and group)
Dense(5, 'sigmoid')	For the categories HS (religion, race, physical, gender, and others)
Dense(3, 'softmax')	For the levels HS (weak, moderate, and strong)

2.6 Performance metrics

To demonstrate the effectiveness of our proposal, we assessed the model's performance using the following metrics: F1-Score, which calculates the harmonic mean of precision and recall, providing insights into testing accuracy. Mathematically expressed as Eq. (2):

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (2)$$

Accuracy: measures the percentage of correct predictions relative to the total number of samples. Can be expressed as Eq. (3):

$$Accuracy = \frac{TP+FN}{TP+FN+TN+FP} \quad (3)$$

Recall: measures the proportion of predicted data in its class. Mathematically, this can be expressed as Eq. (4):

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Precision: measures the likelihood of correctly detected instances of actual events. Calculated as Eq. (5):

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

where, TP, FN, TN, FP correspond to true positive, false negative, true negative, and false positive, respectively.

In addition to evaluating the model using F1-score and label accuracy values, we also utilized the AUC-ROC metric to assess the model's ability to distinguish between positive and negative classes. This process involves calculating the area under the ROC curve, depicting the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at various prediction thresholds [36]. The AUC-ROC metric is employed to evaluate the performance of the classification model by comparing sensitivity (True Positive Rate) and specificity (False Positive Rate). Unlike the confusion matrix, which provides detailed information about the types of errors made by the model, such as errors in predicting positive or negative, the AUC-ROC metric and confusion matrix are often used together to provide a more comprehensive understanding of the classification model's performance. TPR and FPR calculations can be performed using Eq. (6) and Eq. (7).

$$TPR = \frac{TP}{(TP+FN)} \quad (6)$$

$$FPR = \frac{FP}{(FP+TN)} \quad (7)$$

where, TP is the number of True Positives, FN is the number of False Negatives, FP is the number of False Positives, and TN is the number of True Negatives.

3. RESULTS AND DISCUSSION

In order to identify optimal models, we conducted various experimental scenarios and fine-tuned hyperparameters to minimize the loss function for each model. The evaluation of classes involved metrics such as recall, precision, F1 score, and the Area under the ROC Curve (AUC-ROC). These metrics were applied within the context of multi-label classification tasks. The experimental results indicate that the best hyperparameter settings for detecting hate speech and its categories in semantically expanded data include a batch size of 16, a learning rate of 0.001, with epochs set at 20 and 30. On the other hand, for data without semantic expansion, distinct parameters were utilized for each model: the CNN model had a batch size of 16, a learning rate of 1e-04, and 20 epochs; the BiGRU model had a batch size of 16, a learning rate of 3e-05, and 30 epochs; and the BiLSTM model had a batch size of 16, a learning rate of 5e-05, and 30 epochs.

3.1 Results

The obtained results from the performance measurements of our models, derived from experiments involving semantic expansion and non-expanded semantic data in multi-label hate speech, are presented in Figure 4 and Table 4. Consistent parameter settings were maintained across all datasets to ensure optimal model performance. The best parameters, determined through additional experimental testing, were subsequently utilized for training and testing across BiGRU, CNN, and BiLSTM classifications. Our deep learning classifications were tested using three BERT-based embeddings: IndoBERT, IndoBERT-Tweet, and BERT-Multilingual.

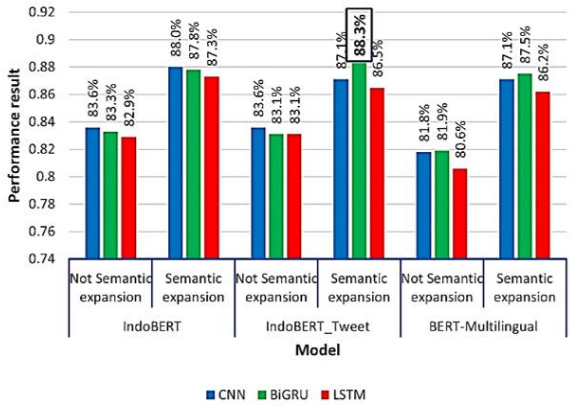


Figure 4. Performance results based on hyperparameter tuning

Table 4. Performance comparison of precision (P), recall (R), and F1-score (%)

Model Embedding	Experiment with	CNN			BiGRU			LSTM		
		P	R	F1	P	R	F1	P	R	F1
IndoBERT	Not semantic expansion	57	69	63	57	69	62	56	69	62
	Semantic expansion	67	80	73	67	81	73	66	78	71
IndoBERT_tweet	Not semantic expansion	58	67	62	57	64	60	57	67	62
	Semantic expansion	65	79	71	68	80	74	64	77	70
BERT-multilingual	Not semantic expansion	52	63	57	49	51	54	47	51	53
	Semantic expansion	66	77	71	66	79	72	64	76	69

Our testing and observations consistently demonstrated that the semantic expansion process significantly improved classification results. This test conclusively indicates that semantic expansion empowers the sentence context to effectively capture scattered and ambiguous word information. Additionally, employing pre-trained BERT-based models during word ranking ensures that only words with the highest scores are included in the semantic expansion process. It's noteworthy that, to maintain conciseness, the addition of new words for semantic expansion was limited to only three words appended to the original sentence.

Figure 4 shows that the embedding model using IndoBERT-Tweet excels in handling semantic features compared to other embedding models. This superiority can be attributed to the fact that the words in the sentences are sourced from Twitter data, although other embedding models also achieve performance levels close to it. When evaluating the classification performance of BiGRU with semantic expansion, it surpasses other classifications such as CNN and LSTM, achieving an accuracy of 88.3%. This represents a better improvement of 0.3% compared to CNN and 1% compared to BiLSTM. In contrast to the performance without semantic expansion, which results in an accuracy of 83.6%, meaning a margin difference of 4.7% compared to semantic expansion performance.

In our performance evaluation, based on precision (P), recall (R), and F1-Score measurements in Table 4, we observe that the application of semantic expansion for BiGRU, LSTM, and CNN classification yields superior performance compared to without semantic expansion. Embedding models based on pre-trained IndoBERT-Tweet, IndoBERT, and BERT-Multilingual for semantic expansion with the BiGRU model generally experience an increase in precision by 2%, recall by 3%, and F1-Score by 2% when handling hate speech texts. Conversely, the process without semantic expansion results in poorer performance for most models. This can be attributed to the involvement of knowledge bases such as WordNet and Kateglo in semantic expansion, enhancing the understanding of words in a broader context. Additionally, pre-trained embedding models like IndoBERT have undergone training on extensive language corpora, providing them with linguistic and semantic knowledge that enables a deeper understanding of word relationships in a broader sentence context.

The experimental results indicate that the optimal hyperparameter settings for detecting hate speech and its categories with semantic expansion data are batch_size=16, LR=0.001, with epochs=20. However, contextual understanding becomes crucial in comprehending the actual impact of these settings. Semantic expansion not only affects hyperparameter setting outcomes but also enriches the meaning of words in sentence contexts. For example, in semantic expansion data, the model produces text representations that encompass more nuanced meanings, thanks to the addition of synonyms, antonyms, and other semantic information. In other words, the model can better capture sentence contexts that can lead to a better understanding of hate speech. Therefore, optimal hyperparameter settings can be viewed as a result of enhanced contextual understanding, leading to better hate speech detection performance.

Conversely, in data without semantic expansion, our models face challenges in capturing deeper nuances in sentences. Different hyperparameter settings are required to ensure optimal performance in this environment. For instance, the CNN model requires a lower learning rate (LR=1e-04) and a shorter number of epochs (epochs=20), while the BiGRU and BiLSTM models have different settings. In data without semantic expansion, we observe that although the models provide good results, there is a significant performance difference compared to data enriched with semantic expansion. This suggests that the absence of additional semantic information can affect the model's ability to understand sentence contexts deeply. Despite achieving satisfactory accuracy levels, further discussion on the constraints and advantages of using semantic expansion compared to data without semantic expansion can provide deeper insights into the concept of contextual understanding. This creates a more comprehensive understanding of how our models perform in various hate speech data contexts.

3.2 Discussion

To address challenges associated with limited data and the absence of semantic features, which can lead to a decline in classification performance, we implemented semantic expansion as an effective solution. In this study, we employed a multi-label hate speech dataset marked by data imbalance. However, through the incorporation of semantic expansion, we significantly enhanced the classification performance compared to the approach without semantic expansion. Our deep learning classification, coupled with pre-trained BERT-based embeddings, demonstrated remarkable effectiveness. In the task of detecting hate speech in Twitter data, we achieved an impressive accuracy rate of 88.3% with the inclusion of semantic expansion. Notably, we observed a substantial increase in recall, showing a remarkable improvement of 30%, along with precision at 21% and an F1-score at 21%, in comparison to the performance without semantic expansion and with semantic expansion.

We would like to highlight the results obtained from the dataset without semantic expansion, where the overall performance in predicting hate speech appears suboptimal. For instance, in the IndoBERT embedding model, CNN, BiGRU, and BiLSTM classifications achieved the highest performance at 83.6%, 83.3%, and 83.1%, respectively. These performance levels were attained through hyperparameter tuning involving variations in epoch and learning rate. For example, CNN performed optimally at LR=1e-04 with an epoch setting of 20, while BiLSTM reached its peak performance at LR=5e-05 with an epoch setting of 30.

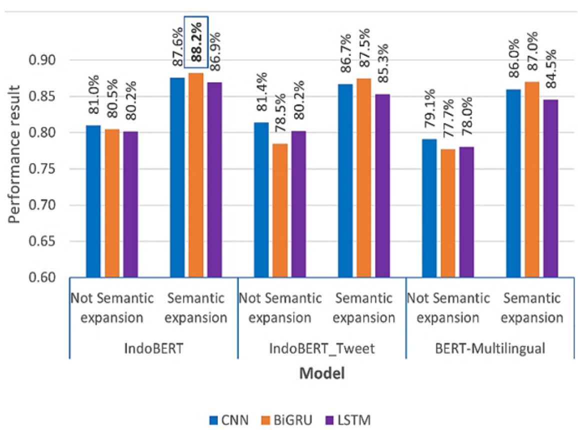


Figure 5. Comparison of evaluation metrics using AUC-ROC curve

When examining the performance metrics of the IndoBERT embedding model, we observe a precision of 67%, recall of 69%, and an F1-score of 63%. Meanwhile, to assess the model's ability to effectively distinguish between positive and negative classes, we employ the AUC-ROC evaluation metric. As depicted in Figure 5, it becomes evident that semantic expansion significantly enhances the model's capacity to differentiate between classes in the classification task. Models incorporating semantic expansion, especially in classification, exhibit substantially higher AUC-ROC values compared to those without semantic expansion, with a notable difference of 6.8%. This observation underscores the superior effectiveness of semantic expansion in understanding the meaning of text and identifying relevant patterns in hate speech data.

The increased AUC-ROC values in models with semantic expansion can be explained by the addition of extra semantic information that enriches word representations, aiding the model in comprehending complex contexts and variations in meaning. By incorporating synonyms, antonyms, and other semantic information, the model becomes more sensitive to meaning variations, enhancing its ability to distinguish between positive and negative classes. Improved word representations also provide the model an advantage in handling variations in word meaning, enabling it to adapt better to complex test data. Overall, semantic expansion provides valuable additional information to the model, enabling it to perform better in classification tasks, as reflected in the increased AUC-ROC values.

To facilitate a comprehensive comparison, we present four prior studies that employed multi-label and multi-output datasets in the Indonesian language, as outlined in Table 5. The rationale for choosing these earlier studies as benchmarks is to assess the effectiveness of our approach, both with and without semantic expansion, on Indonesian text using the identical dataset. It's noteworthy that some of these previous studies employed all available labels, while others concentrated on specific subsets during their experiments.

Table 5. Performance of studies on hate speech detection

Method	Avg. acc (%)
BiGRU+IndoBERT [37]	83.8
SVM [38]	68.3
BiLSTM+BERT [39]	64.8
RFDT+CC	76.1
CNN+DistilBERT [40]	61.3
SVM+CC [40]	74.8
Proposed	
SemEx+IndoBERT+CNN	88.0
SemEx+IndoBERT_tweet+BiGRU	88.3
SemEx+IndoBERT+BiLSTM	87.3

In our investigation, we systematically assessed the effectiveness of the proposed semantic expansion approach. Previous studies have employed a variety of methods, ranging from machine learning-based approaches to deep learning techniques, such as SVM, Random Forest Decision Tree (RFDT), Naive Bayes (NB), CNN, LSTM, BiLSTM, and BiGRU, to evaluate hate speech data in the context of Indonesia. Despite the diverse range of classification techniques used, all of them demonstrated limited effectiveness in hate speech detection. Additionally, these studies incorporated various pre-processing techniques. For example, Prabowo et al. [38] utilized the SVM algorithm and uni-gram word features, selecting nine labels out of a total of twelve, achieving an accuracy rate of 68.42%. Subsequently, in Marpaung et al. [37], the BiGRU+Word2vec model was implemented, demonstrating the best performance at 76.70%, which further improved to 84.77% with the IndoBERT+BiGRU model. Taking a different approach, Hendrawan et al. [39] and Hana et al. [40] employed translation methods and data transformation using Classifier Chains (CC) on the entire dataset. However, this approach proved effective only in the SVM model, with the highest accuracy reaching 74.88%, whereas in the RFDT model, the accuracy increased to 76.12%. Notably, they also conducted experiments with CNN and BiLSTM models, yielding CNN+BERT performance at 61.30% and BiLSTM+BERT at 64.81%.

While our experiments delivered more promising results, surpassing previous research, our dataset posed challenges related to label imbalance (Table 1), leading to inconsistencies in precision, recall, F1-score, and accuracy metrics. Our study exclusively focused on testing using the semantic expansion approach. Despite exploring various class-balancing techniques, such as applying class-weighting, these attempts were unsuccessful due to the lack of support for this technique in the latest TensorFlow version.

As a result, among the various experimental scenarios investigated, we exclusively employed techniques like back-translation and semantic expansion (SemEx) to enhance classification performance. Throughout the classification process, our emphasis was solely on implementing hyperparameter tuning to ensure the most effective configuration. It's crucial to note that the back-translation technique introduces changes to sentence structure, influencing contextual understanding. Therefore, this phase can be adjusted to meet specific requirements in alignment with the chosen approach and the data domain in use.

4. CONCLUSION AND FUTURE WORK

The main objective of this research is to generate concise text representations using a semantic expansion approach and overcome challenges related to insufficient data and the absence of semantic features in conventional methods, especially in the context of hate speech data. To achieve this, we propose a hybrid semantic expansion method that combines text expansion techniques utilizing the knowledge bases WordNet and Kateglo, along with word ranking based on semantic similarity. Applying semantic expansion to multi-label and multi-output datasets presents a notable challenge for classification algorithms dealing with data imbalance in hate speech. However, based on experimental results, we confirm that using semantic expansion significantly enhances the performance of classification algorithms such as CNN, BiGRU, and BiLSTM. The performance of these algorithms markedly improves compared to the approach without semantic expansion for hate speech detection.

In terms of implications, this research has a substantial impact on advancing hate speech detection techniques. The experimental findings, demonstrating improved performance of classification algorithms with the hybrid semantic expansion approach, offer valuable insights for researchers and practitioners aiming to enhance the accuracy of hate speech detection. Additionally, this study may aid in the development of other models relying on short text representations. Moreover, it contributes to a better understanding of how technology can be effectively employed to detect hate speech, positively influencing security and comfort in the online environment. The potential improvement in hate speech detection holds the promise of creating a safer and more positive online environment for users in general.

Table 6. Real-time data testing

Sentence	Classifier Model	Label-BASED HATE SPEECH (HS) Detection		
		Not Semantic Expansion	Semantic Expansion (Using Best Models)	Using Best Model and Text Expansion
The infidel tadpoles have looked stupid from the start and are even more stupid hahaha.	CNN	HS (70%), Processing time=4 sec.	HS (84%), Processing time=8 sec.	HS (89%), Processing time=1 minute 15 sec.
	BiGRU	HS (66%), Processing time=2 sec.	HS (93%), Processing time=3 sec.	HS (98%), Processing time=51 sec.
	BiLSTM	HS (79%), Processing time=2 sec.	HS (84%), Processing time=4 sec.	HS (94%), Processing time=52 second.
People with devil hearts like this don't need to be forgiven.	CNN	HS (70%), Processing time=4 sec.	HS (84%), Processing time=8 sec.	HS (89%), Processing time=1 minute 15 sec.
	BiGRU	HS (66%), Processing time=2 sec.	HS (93%), Processing time=3 sec.	HS (98%), Processing time=51 second.
	BiLSTM	HS (79%), Processing time=2 sec.	HS (84%), Processing time=4 sec.	HS (94%), Processing time=52 sec.

While our research has made a significant contribution to the field of natural language processing through semantic expansion for deep learning classification tasks, especially concerning imbalanced multi-label and multi-output hate speech data, there are still limitations that warrant further investigation. Firstly, we observed that the improvement in performance depends on the number of word samples used in the expansion process, impacting computational time. Our approach requires extended computational time compared to methods without semantic expansion (Table 6). This is attributed to the text expansion process, involving the extraction of synsets or sets of synonymous words from WordNet and Kateglo. In contrast, methods without semantic expansion do not involve retrieving synonym sets from external knowledge bases. Secondly, determining the optimal model through hyperparameter tuning necessitates substantial computational resources, proving impractical for the numerous parameter settings in our experiments. We conducted hyperparameter tuning separately for the IndoBERT model to conserve resources, thereby prolonging the duration of the process.

REFERENCES

[1] Schmidt, A., Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1-10. https://doi.org/10.18653/v1/W17-1101

[2] Manusia, K.N.H.A. (2015). Buku saku penanganan ujaran kebencian (hate speech). Jakarta: Komnas HAM. http://ylbhu.org/wp-content/uploads/2016/06/ Panduan-Teoritik-Hate-Speech.pdf.

[3] Naseem, U., Razzak, I., Eklund, P.W. (2021). A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on twitter. Multimedia Tools and Applications, 80: 35239-35266. https://doi.org/10.1007/s11042-020-10082-6

[4] Cheng, X., Zhang, C., Li, Q.X. (2021). Improved Chinese short text classification method based on ERNIE_BiGRU model. In Journal of Physics: Conference Series, IOP Publishing, 1993(1): 012038. https://doi.org/10.1088/1742-6596/1993/1/012038

[5] Wang, P., Xu, B., Xu, J.M., Tian, G.H., Liu, C.L., Hao, H.W. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174: 806-814. https://doi.org/10.1016/j.neucom.2015.09.096

[6] Wang, H.B., Luo, G.R., Li, R.X. (2021). A short text classification method based on combining label information and self-attention graph convolutional neural network. In Computer Supported Cooperative Work and Social Computing: 15th CCF Conference, ChineseCSCW, Springer Singapore, pp. 670-677. https://doi.org/10.1007/978-981-16-2540-4_50

[7] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv Preprint arXiv: 1301.3781. https://doi.org/10.48550/arXiv.1301.3781

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint arXiv: 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[9] Chen, Y., Zhang, H., Liu, R., Ye, Z.W., Lin, J.Y. (2019). Experimental explorations on short text topic mining between LDA and NMF based schemes. Knowledge-Based Systems, 163: 1-13. https://doi.org/10.1016/j.knosys.2018.08.011

[10] Jelodar, H., Wang, Y.L., Yuan, C., Feng, X., Jiang, X.H., Li, Y.C., Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78: 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

[11] Naskar, S.K., Bandyopadhyay, S. (2007). Word sense disambiguation using extended wordnet. In 2007 International Conference on Computing: Theory and Applications (ICCTA'07), IEEE, pp. 446-450. https://doi.org/10.1109/ICCTA.2007.134

[12] Chaplot, D.S., Salakhutdinov, R. (2018). Knowledge-based word sense disambiguation using topic models. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12027

[13] Mahmoud, A., Zrigui, M. (2017). Semantic similarity analysis for paraphrase identification in Arabic texts. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, pp. 274-281.

[14] Lu, M.L., Sun, X.B., Wang, S.W., Lo, D., Duan, Y.C. (2015). Query expansion via wordnet for effective code search. In 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER), pp. 545-549. https://doi.org/10.1109/SANER.2015.7081874

[15] Azad, H.K., Deepak, A., Chakraborty, C., Abhishek, K.

(2022). Improving query expansion using pseudo-relevant web knowledge for information retrieval. Pattern Recognition Letters, 158: 148-156. https://doi.org/10.1016/j.patrec.2022.04.013

[16] Muzakir, A., Adi, K., Kusumaningrum, R. (2023). Advancements in semantic expansion techniques for short text classification and hate speech detection. Ingenierie des Systemes d'Information, 28(3): 545-556. https://doi.org/10.18280/isi.280302

[17] Wang, P., Xu, J.M., Xu, B., Liu, C.L., Hao, H.W. (2015). A convolutional architecture for short text expansion and classification. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 1: 75-78. https://doi.org/10.1109/WI-IAT.2015.12

[18] Adhi, M.S., Nafan, M.Z., Usada, E. (2019). Pengaruh semantic expansion pada naïve bayes classifier untuk analisis sentimen tokoh masyarakat. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 3(2): 141-147. https://doi.org/10.29207/resti.v3i2.901

[19] Jahan, M.S., Beddiar, D.R., Oussalah, M., Mohamed, M. (2022). Data expansion using wordnet-based semantic expansion and word disambiguation for cyberbullying detection. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 1761-1770.

[20] Minh, D.L., Sadeghi-Niaraki, A., Huy, H.D., Min, K., Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access, 6: 55392-55404. https://doi.org/10.1109/ACCESS.2018.2868970

[21] Zulqarnain, M., Abd Ishak, S., Ghazali, R., Nawi, N.M., Aamir, M., Hassim, Y.M.M. (2020). An improved deep learning approach based on variant two-state gated recurrent unit and word embeddings for sentiment classification. International Journal of Advanced Computer Science and Applications, 11(1): 594-603. https://doi.org/10.14569/ijacsa.2020.0110174

[22] Ibrohim, M.O., Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. In Proceedings of the Third Workshop on Abusive Language Online, pp. 46-57. https://doi.org/10.18653/v1/w19-3506

[23] Sastrawan, I.K., Bayupati, I.P.A., Arsa, D.M.S. (2022). Detection of fake news using deep learning CNN-RNN based methods. ICT Express, 8(3): 396-408. https://doi.org/10.1016/j.icte.2021.10.003

[24] Laatar, R., Aloulou, C., Belghuith, L.H. (2018). Word embedding for Arabic word sense disambiguation to create a historical dictionary for Arabic language. In 2018 8th International Conference on Computer Science and Information Technology (CSIT), IEEE, pp. 131-135. https://doi.org/10.1109/CSIT.2018.8486159

[25] Le, Q., Mikolov, T. (2014). Distributed representations of sentences and documents. In International Conference on Machine Learning, PMLR, 32(2): 1188-1196. https://doi.org/10.48550/arXiv.1405.4053

[26] Miao, Y.L., Ji, Y.C., Peng, E. (2019). Application of CNN-BiGRU model in Chinese short text sentiment analysis. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, pp. 510-514. https://doi.org/10.1145/3377713.3377804

[27] Koto, F., Lau, J.H., Baldwin, T. (2021). IndoBERTweet: a pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10660-10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

[28] Kulmanov, M., Smaili, F.Z., Gao, X., Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. Briefings in Bioinformatics, 22(4): bbaa199. https://doi.org/10.1093/bib/bbaa199

[29] Harjo, B., Muljono, Abdullah, R. (2023). Attention-based sentence extraction for aspect-based sentiment analysis with implicit aspect cases in hotel review using machine learning algorithm, semantic similarity, and BERT. International Journal of Intelligent Engineering & Systems, 16(3): 189-200. https://doi.org/10.22266/ijies2023.0630.15

[30] Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X.H., Lim, Z.Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. arXiv Preprint arXiv: 2009.05387. https://doi.org/10.48550/arXiv.2009.05387

[31] Rochmawati, N., Hidayati, H.B., Yamasari, Y., Tjahyaningtijas, H.P.A., Yustanti, W., Prihanto, A. (2021). Analisa learning rate dan batch size pada klasifikasi covid menggunakan deep learning dengan optimizer adam. Journal Information Engineering and Educational Technology, 5(2): 44-48. https://doi.org/10.26740/jieet.v5n2.p44-48

[32] Chowanda, A., Chowanda, A.D. (2017). Recurrent neural network to deep learn conversation in indonesian. Procedia Computer Science, 116: 579-586. https://doi.org/10.1016/j.procs.2017.10.078

[33] Zhou, Y.J., Xu, J.M., Cao, J., Xu, B., Li, C.L. (2017). Hybrid attention networks for Chinese short text classification. Computación y Sistemas, 21(4): 759-769. https://doi.org/10.13053/cys-21-4-2847

[34] Qiu, H., Fan, C.D., Yao, J., Ye, X.H. (2020). Chinese microblog sentiment detection based on CNN-BiGRU and multihead attention mechanism. Scientific Programming, 2020: 1-13. https://doi.org/10.1155/2020/8865983

[35] Ependi, U., Rochim, A.F., Wibowo, A. (2023). A hybrid sampling approach for improving the classification of imbalanced data using ROS and NCL methods. International Journal of Intelligent Engineering and Systems, 16(3): 345-361. https://doi.org/10.22266/ijies2023.0630.28

[36] Ayo, F.E., Folorunso, O., Ibharalu, F.T., Osinuga, I.A., Abayomi-Alli, A. (2021). A probabilistic clustering model for hate speech classification in twitter. Expert Systems with Applications, 173: 114762. https://doi.org/10.1016/j.eswa.2021.114762

[37] Marpaung, A., Rismala, R., Nurrahmi, H. (2021). Hate speech detection in Indonesian Twitter texts using bidirectional gated recurrent unit. In 2021 13th International Conference on Knowledge and Smart Technology (KST), IEEE, pp. 186-190. https://doi.org/10.1109/KST51265.2021.9415760

[38] Prabowo, F.A., Ibrohim, M.O., Budi, I. (2019). Hierarchical multi-label classification to identify hate speech and abusive language on Indonesian twitter. In 2019 6th International Conference on Information Technology, Computer and Electrical Engineering

(ICITACEE) IEEE, pp. 1-5. https://doi.org/10.1109/ICITACEE.2019.8904425
[39] Hendrawan, R., Adiwijaya, Al Faraby, S. (2020). Multilabel classification of hate speech and abusive words on Indonesian Twitter social media. In 2020 International Conference on Data Science and Its Applications (ICoDSA), IEEE, pp. 1-7. https://doi.org/10.1109/ICoDSA50139.2020.9212962
[40] Hana, K.M., Adiwijaya, Al Faraby, S., Bramantoro, A. (2020). Multi-label classification of indonesian hate speech on twitter using support vector machines. In 2020 International Conference on Data Science and Its Applications (ICoDSA), IEEE, pp. 1-7. https://doi.org/10.1109/ICoDSA50139.2020.9212992