International Journal of Safety and Security Engineering
Vol. 15, No. 3, March, 2025, pp. 509-520
Journal homepage: http://iieta.org/journals/ijsse

A Hybrid Semantic Enrichment Approach for Multi-Label Toxic Speech Detection

Ari Muzakir[1,2*], Uci Suriani[3], Usman Ependi[2]

[1] Faculty of Science and Technology, Bina Darma University, Palembang 30111, Indonesia
[2] Master of Informatics Engineering Study Program, Bina Darma University, Palembang 30111, Indonesia
[3] Multimedia Engineering Technology Study Program, Prasetiya Mandiri Darussalam Polytechnic, Palembang 30128, Indonesia

Corresponding Author Email: arimuzakir@binadarma.ac.id

https://doi.org/10.18280/ijsse.150310

Received: 11 February 2025
Revised: 18 March 2025
Accepted: 22 March 2025
Available online: 31 March 2025

Keywords:
deep learning for text classification, multi-label classification, semantic enrichment in NLP, social media hate speech, toxic speech detection

ABSTRACT

The rapid growth of digital communication has facilitated the spread of toxic speech, which can harm individuals or communities and often appears across multiple nuanced categories. These categories are difficult to detect in short texts due to semantic ambiguity, limited context, and label dependencies. This study introduces a Hybrid Semantic Enrichment with Convolutional Neural Network (HSE-CNN) approach to enhance multi-label toxic speech classification. The HSE-CNN model leverages semantic enrichment techniques such as back translation, text expansion, word sense disambiguation (WSD), and semantic similarity mapping to enrich the contextual meaning of input texts. Using an Indonesian social media dataset containing 13,169 entries labeled with 12 toxic speech categories, we conducted a series of experiments involving preprocessing, semantic enrichment, and classification using various deep learning models. The optimal configuration includes a learning rate of 0.001, batch size of 16, and training for 30 epochs. Our proposed model achieved an F1-score of 80%, accuracy of 93%, and AUC of 91%, demonstrating its superiority over non-enriched models. Compared to baseline models such as BiLSTM and BiGRU, the HSE-CNN yields a 6.7% improvement in accuracy and a 4.5% improvement in F1-score. These findings suggest that HSE-CNN offers a promising solution for toxic speech detection systems, especially in resource-limited languages, with potential applications in digital content moderation, online safety initiatives, and public awareness enhancement.

1. INTRODUCTION

The rapid advancement of digital technology has significantly facilitated the production and dissemination of user-generated content, making it faster, more accessible, and cost-effective. Alongside these advantages, this development has also contributed to the widespread proliferation of toxic speech, which can be easily shared across social media platforms with minimal effort and often under the guise of anonymity. The uncontrolled spread of harmful content poses serious risks, including escalating social conflicts, fostering discrimination, and inciting real-world harm [1]. Due to these risks, effective mechanisms to regulate and detect toxic speech have become an urgent necessity.

Among various social media platforms, Twitter (X) serves as a major space for real-time public discourse, allowing users to freely share opinions, criticisms, and emotions. While this openness fosters engagement, it also makes X media social a potential hotspot for toxic speech [2]. The ability to detect and mitigate toxic speech is crucial for preventing cybercrimes, minimizing harm in online communities, and fostering a healthier digital environment [3, 4]. However, identifying toxic speech is particularly challenging due to its diverse expressions, contextual dependencies, and linguistic variations. Toxic speech is often multi-dimensional, involving multiple overlapping categories such as targeting specific individuals or groups, addressing sensitive topics like race, religion, or gender, and varying in intensity from mild to severe hostility [5, 6].

One of the key challenges in toxic speech detection lies in its semantic complexity. Toxic expressions are not always overt but can be conveyed through sarcasm, implicit meaning, metaphors, and coded language, making detection difficult. Moreover, the brevity of social media posts, especially on platforms like X media social that impose character limits, further complicates the ability of models to capture the necessary context [7, 8]. Traditional methods such as Latent Semantic Analysis (LSA) have been employed to extract patterns from short texts [9], but these methods face notable limitations in handling synonyms, understanding word order, and maintaining contextual coherence [10]. Recent studies have attempted to improve these limitations through contextualized variants of LSA, such as LSA-BERT [11], and LSA with transformer-based embeddings [12, 13], yet their performance in dynamic social media environments remains limited.

To improve toxic speech classification, various approaches have been proposed, including knowledge-based semantic

enrichment techniques and word embedding representations. Semantic enrichment methods have emerged as more effective solutions for resolving word ambiguity [14] and enhancing semantic comprehension [15]. These methods improve semantic analysis by integrating knowledge-based approaches, word embedding representations, and hybrid techniques [16, 17]. Knowledge-based methods, such as WordNet-based thesauruses, ConceptNet-based ontologies, and specialized lexicons like Kateglo, provide structured linguistic knowledge to resolve word ambiguities and improve interpretation accuracy [12-14]. However, these methods often struggle to handle evolving language trends, newly introduced slang, and domain-specific terminology [18]. Recent evaluations have also shown that rigid ontologies and rule-based knowledge systems lack flexibility when applied to informal, multilingual, and sarcastic content in toxic speech detection tasks [19-21].

On the other hand, word embedding techniques like Word2Vec, GloVe, and BERT have been widely used to model semantic relationships between words [16, 17]. While these methods capture contextual similarities, they often lack explicit semantic structures, making them prone to misclassification in cases of subtle meaning variations. For example, embedding-based models may assign high similarity scores to semantically unrelated terms due to shared contexts, introducing noise in the classification process [22].

To overcome semantic analysis limitations such as contextual bias, semantic restrictions, and challenges in multi-label classification, hybrid approaches integrate multiple methodologies to achieve a more comprehensive understanding. These methods have demonstrated effectiveness in enhancing semantic analysis performance [23] and improving classification accuracy [24], particularly for multi-label datasets. Classification accuracy serves as a critical metric in evaluating machine learning models. Traditional machine learning techniques often struggle to capture the complexities of label dependencies in toxic speech classification. For instance, models such as Support Vector Machine (SVM) and Random Forest Decision Tree (RFDT) rely heavily on manually engineered features and are generally less effective in capturing intricate label relationships. Even deep learning approaches such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks face challenges as they process data sequentially and lack the parallelization capabilities necessary for real-time toxic speech detection [25]. Recent benchmarks have shown that LSTM-based models still suffer from latency issues and underperform compared to CNNs [26]. LSTM-based processes data sequentially but is not effective in parallel, while CNN can process data in parallel and is faster in training [6].

By addressing the existing challenges in toxic speech classification, this study introduces a hybrid feature extraction approach based on Semantic Enrichment and Convolutional Neural Networks (CNN) to improve multi-label toxic speech classification. The model integrates various natural language processing (NLP) techniques such as back translation, word sense disambiguation, text expansion, and semantic similarity measurement to provide deeper contextual understanding and richer feature representations. By enhancing input texts with additional semantic context, this approach improves classification accuracy across multiple labels, significantly enhancing model performance compared to traditional approaches.

This study makes several important contributions. First, it proposes a novel hybrid feature extraction model that leverages semantic enrichment techniques to enhance multi-label classification performance. By integrating structured linguistic knowledge with deep learning architectures, the model improves the ability to detect contextual variations and complex toxic speech patterns [27]. Second, the study demonstrates that the HSE-CNN approach outperforms traditional models such as SVM, RFDT, BiGRU, and BiLSTM, particularly in addressing challenges related to ambiguity, brevity, and multi-label dependencies in toxic speech classification. Third, this research provides a detailed analysis of the impact of each semantic enrichment technique, offering insights into how different linguistic enhancements contribute to model performance. These findings are valuable for advancing AI-driven content moderation systems and improving the reliability of automated toxic speech detection in social media environments.

By tackling the inherent challenges in toxic speech classification, this study contributes to the development of robust, context-aware models that can be applied in various content moderation applications, policy enforcement strategies, and social media monitoring tools. The integration of semantic enrichment with deep learning represents a promising direction for enhancing the interpretability and effectiveness of AI-based toxic speech detection systems.

2. RELATED WORKS

2.1 Multi-label text classification

Traditional classification models are typically designed to assign a single label to each sample in datasets with single-label annotations. However, in more complex tasks such as toxic speech classification, models need to predict multiple labels simultaneously for each sample. This introduces a challenge, as multi-label classification often results in lower performance compared to single-label scenarios. For instance, Alfina et al. [28] achieved an accuracy of 92.5% using a Random Forest Decision Tree (RFDT) model on single-label data to differentiate between neutral and toxic speeches. In contrast, Hendrawan and Al Faraby [29], using RFDT for multi-label toxic speech classification with 12 labels, obtained a significantly lower accuracy of 76.1%. This discrepancy stems from the inherent complexity of predicting multiple labels, which is more challenging than handling single-label data samples.

To address these challenges, several researchers have explored data transformation and algorithm adaptation strategies [30]. For example, Ibrohim and Budi [31], applied techniques such as Classifier Chains (CC), Label Powerset (LP), and Binary Relevance (BR) to enhance multi-label classification performance. Moreover, algorithm adaptation has involved the use of specialized classification algorithms like Multi-label KNN, Multi-Label SVM, and ensemble methods such as ML-RF [30]. Additionally, Prabowo et al. [32] tackled the multi-label classification problem by reducing the number of labels from 12 using various SVM models. These approaches aim to improve classifiers' ability to handle the complexities of multiple labels while mitigating the challenges posed by high-dimensional label spaces.

2.2 Semantic enrichment

Semantic enrichment plays a crucial role in providing a

broader context to short texts. The primary strategy involves identifying relevant words or phrases based on user intent and integrating them into initial outputs, thereby enhancing the model's comprehension. This process aims to mitigate word ambiguity and generate more comprehensive information [33]. One of the key semantic enrichment techniques highlighted in recent studies is data augmentation, also known as back translation [34]. This technique involves translating text into another language and then translating it back, generating a wider range of semantic variations without altering the original meaning. Research by Beddiar et al. [15] and Siino et al. [35], has demonstrated that back translation can significantly improve model performance in tasks such as toxic speech detection, particularly for short and ambiguous texts. These enrichment techniques are particularly valuable in overcoming the challenges posed by text brevity and word ambiguity issues that are common in platforms like X media social, where opinions are often expressed concisely and can be influenced by various nuances.

In addition to back translation, researchers have also turned to diverse knowledge repositories, such as thesauruses, dictionaries, and ontologies, to enrich the semantic content of text [36]. For example, Voorhees [37] developed the lexical resource WordNet to identify synonymous terms, while Azad and Deepak [38] utilized both Wikipedia and WordNet to expand words with weighted terms. This enabled a better understanding of contextual relationships, contributing to a richer vocabulary and more nuanced meaning. Such methods help the model better distinguish between varying levels of toxicity in opinions and detect subtle differences in meaning. This integration of knowledge sources is particularly beneficial in the detection of toxic speech, where the context and interpretation of words can significantly impact their perceived negativity.

2.3 Hybrid classification models

Recent advancements in natural language processing (NLP) have highlighted the significance of integrating semantic enrichment techniques with deep learning models to enhance the performance of multi-label classification tasks. Deep learning models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, are highly effective in capturing local patterns in text, such as n-grams or sequential dependencies [39]. However, these models often struggle with understanding broader contextual relationships, particularly in short, ambiguous, or multi-labeled texts. The challenge arises from the need to balance the efficient capture of local features with the ability to understand global context, especially when text contains multiple categories and varying levels of toxicity.

To address these limitations, several semantic enrichment techniques, such as back translation, word sense disambiguation (WSD), and text expansion have been integrated into deep learning architectures. These techniques enable models to move beyond surface-level syntactic features and capture deeper semantic relationships, thereby improving classification accuracy in multi-label scenarios [40]. By combining CNN with semantic enrichment layers, models can process local text patterns efficiently while also gaining a broader understanding of the global context. This hybrid approach not only enhances feature representation but also makes it particularly effective for tasks like toxic speech classification, where nuanced contextual relationships between multiple labels (e.g., target, category, intensity) are essential for accurate predictions.

Moreover, this hybrid approach facilitates better generalization by allowing the model to focus on both the detailed aspects of the text and the broader context in which these aspects appear. This is especially advantageous for online content platforms like X media social, where language is often informal, context-dependent, and occasionally ambiguous [41]. By integrating semantic enrichment with deep learning models, the classification task becomes more robust, enabling the model to effectively handle the complexity and variability inherent in online toxic speech data.

3. PROPOSED METHOD

In this study, we introduce a Hybrid Semantic Enrichment and Neural Network based on CNN (HSE-CNN) method to enhance semantic understanding and improve the performance of toxic speech classification in multi-label datasets. The methodology involves several preprocessing steps followed by semantic expansion. We employ techniques such as back-translation, word sense disambiguation, text expansion, and semantic similarity to create a robust, hybrid approach. Vector representations are generated using IndoBERT, and deep learning models such as CNN, BiGRU, and BiLSTM are employed for classification to identify the optimal model for toxic speech detection. The overall schematic of this study is illustrated in Figure 1.

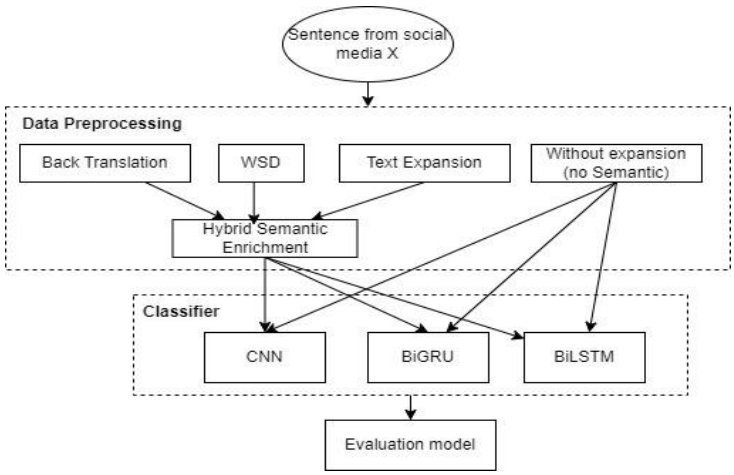

Figure 1. Working model of toxic detection

3.1 Data

The dataset used in this study consists of 13,169 tweets from X media social [31], annotated with 12 multi-label categories relevant to toxic speeches. Thirty linguists conducted the annotation to ensure the validity of the labels. Multi-label classification introduces unique challenges such as class imbalance, high dimensionality, and label interdependencies. For instance, a tweet may simultaneously be labeled as both Abuse and Toxic speech if it contains offensive language or threats. To address these complexities, the labels are organized into hierarchical groupings based on three key aspects: target (individual or group), category (religion, race, physical, gender, and offensive), and intensity level (weak, moderate, or strong). This hierarchical structure provides a systematic and interpretable approach to multi-label annotation.

Multi-label data arises when a sample belongs to more than one class, making classification challenging as the model must recognize multiple relevant labels. This issue is especially pronounced in cases of class imbalance, high dimensionality, and label interactions. For example, a text could be labeled both as Abuse and Toxic speech if it contains offensive language or threats. To handle this, labels are organized hierarchically by target, category, and intensity. Each label is assigned an output layer corresponding to its characteristics, allowing for a comprehensive understanding of the hierarchy of toxic speech labels.

3.2 Hybrid semantic enrichment

To improve text comprehension, this study employs HSE through four stages: back-translation, word sense disambiguation, text expansion, and semantic similarity, which are explained and illustrated in Figure 2.

Based on Figure 2, the HSE process begins with identifying the target word in a sentence, such as "*lit*" in "*That party was lit!*". The target word is then analyzed using the Lesk Algorithm in the word sense disambiguation (WSD) stage to determine the most appropriate synset from lexical sources like WordNet and Kateglo. The disambiguation process identifies relevant synonyms for "*lit*", such as "*amazing*", "*very enjoyable*", and "*great*". Next, a text expansion strategy is applied by incorporating these synonyms into the original sentence, resulting in "*That party was lit, amazing, great, very enjoyable*". To further enhance semantic understanding, each word in the sentence is represented as a vector using BERT embeddings. For example, "*amazing*" has a vector representation of [0.5, 0.7], "*very enjoyable*" [0.3, 0.8], and "*great*" [-0.6, 0.4]. Finally, semantic similarity is measured using Cosine Similarity, which evaluates contextual relationships between words. For instance, ("*amazing*", "*damned*") = 0.95 indicates a strong semantic association, whereas ("*very enjoyable*", "*abominable*") = 0.02 suggests a weak relationship. This process enables NLP models to better understand word context, resolve linguistic ambiguities, and enhance performance in semantic analysis and text classification tasks.

3.2.1 Back-translation

In this stage, the text is translated into a different language and then back to the original to enhance comprehension and refine sentence structure. Using the Google Translate API, the text is first translated into another language and then returned to its original form. This process captures a variety of expressions and nuances, enriching the semantic content of the text. The back-translation follows the equation, as shown in Eq. (1).

$$T' = \textit{Back-translation}\ (T) \tag{1}$$

where, T is the original text, and T' is the resulting back-translated text. For example, "*That party was lit!*" is translated into German as "*Die Party war der Hammer!*" and then back to English as "*The party was the hammer!*", which reveals linguistic variations and deepens semantic understanding.

3.2.2 Word sense disambiguation

Word sense disambiguation is used to resolve ambiguities in word meanings based on context. This is achieved by utilizing WordNet and the Lesk algorithm to identify the correct meaning of a word in a given context, as shown in Eq. (2):

$$\textit{synset}(t,C) = \textit{Lesk}(t,C,\textit{WordNet}) \tag{2}$$

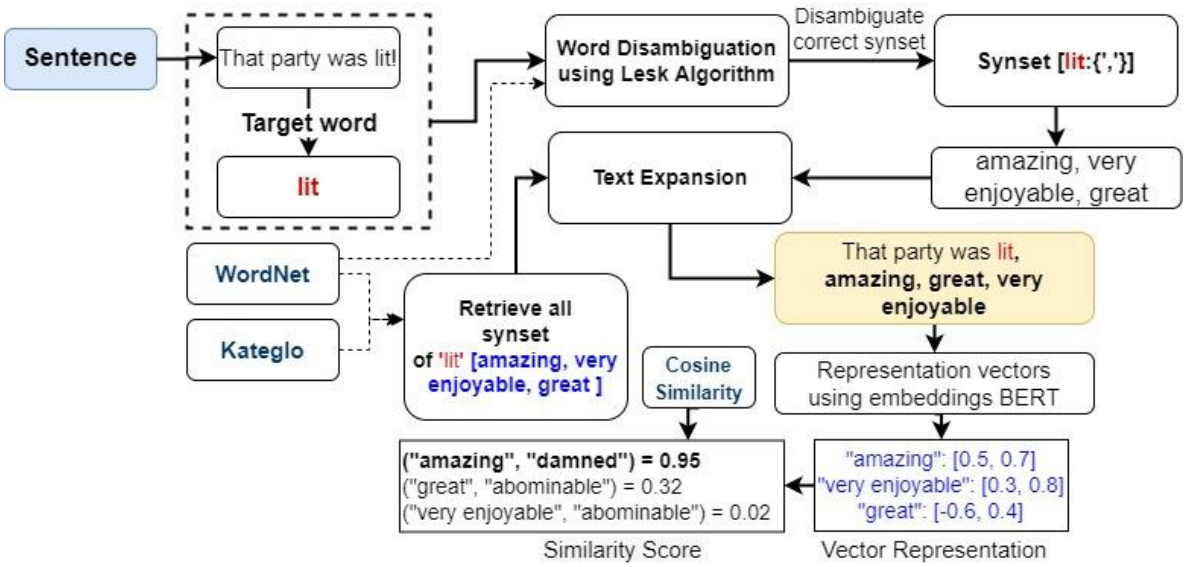

Figure 2. Example of the HSE process

where, *t* represents the word, *C* represents the context, and synset(*t, C*) represents the resulting word sense. This process helps clarify ambiguities, providing more accurate interpretations of words. For example, the word "*lit*" in "*That party was lit!*" could mean "*illuminated*" or "*very enjoyable*." The Lesk algorithm clarifies that in the context of a party, "*lit*" means "*amazing*," yielding a clearer statement: "*That party was amazing!*".

3.2.3 Text expansion

Text expansion is a technique used to enrich the original input text by adding semantically related words to improve contextual understanding. In this study, we expand the input text by appending relevant synonyms to selected keywords. Synonym candidates are initially retrieved from lexical databases such as WordNet and Kateglo. However, to ensure that only contextually appropriate synonyms are included, we implement a two-stage filtering process.

First, we apply a semantic similarity threshold using IndoBERT pre-trained embeddings, retaining only those synonyms whose cosine similarity with the original word exceeds 0.7. This threshold ensures that added words maintain close semantic proximity to the original context.

Second, we apply part-of-speech (POS) tagging and contextual matching. Only synonyms that share the same POS tag and appear in similar syntactic and semantic contexts are selected. This helps avoid noise from grammatically or semantically incompatible additions.

- *Rule 1*: S ← Synonyms(target_word) from {WordNet ∪ Kateglo}.
- *Rule 2*: S' ← {w ∈ S | POS(w) = POS(target_word)}
- *Rule 3*: S" ← {w ∈ S' | cosine_similarity(IndoBERT(w), IndoBERT(target_word)) ≥ 0.7}
- *Rule 4*: S_final ← {w ∈ S" | context_match(w, context_window) = TRUE}

Example: Synonym Selection Process for the word *"berbahaya* (dangerous)":

- Step 1: Retrieve candidate synonyms from lexical databases such as WordNet Bahasa and Kateglo. Result: ["*mengancam* (threatening)", "*mematikan* (deadly)", "*menyeramkan* (scary)", "*tidak aman* (unsafe)"].
- Step 2: Filter the retrieved candidates to include only those that match the target word's POS (e.g., adjectives). Result after filtering: ["threatening ", " deadly ", " scary "].
- Step 3: Compute cosine similarity using pre-trained IndoBERT embeddings. Only synonyms with a similarity score ≥ 0.70 are retained.
  1. " threatening " → 0.82
  2. " deadly " → 0.79
  3. " scary " → 0.65
- Step 4: The final set of synonyms used for text expansion: ["threatening ", " deadly "].

3.2.4 Semantic similarity

To evaluate the semantic similarity between the original and expanded texts, we use cosine similarity and pre-trained BERT. This ensures that the expanded text maintains semantic coherence. Semantic similarity between words or phrases in the text is computed using Eq. (3).

$$Sim(v1, v2) = \frac{v1.v2}{||v1||||v2||} \quad (3)$$

where, *v1* and *v2* are embedding vectors of two words generated by BERT.

3.3 Classification model

The toxic speech classification process consists of two main stages: data partitioning and model development. Data partitioning involves five types of preprocessed and semantically enriched data: data without semantic enhancements, back-translation data, text expansion and disambiguation data, text expansion data alone, and hybrid semantic enrichment data. The data is split into 80% for training and validation, and 20% for testing, using k-fold cross-validation with k = 5 to optimize data usage. Model development and evaluation involve the use of IndoBERT embeddings and a CNN-based text classification model. This model is compared with others, including Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM), to assess their effectiveness in toxic speech classification.

The text classification model utilized is a 1D CNN, known for its effectiveness in text classification tasks. The process involves constructing and training the model, followed by validation and testing. Hyperparameters such as the number of convolutional layers, kernel sizes, and activation functions are tuned for optimal model performance. Validation is carried out using parameters such as Batch Size, Epoch, and Learning Rate.

The CNN architecture includes several layers: input, embedding, Convolution Layer 1, pooling layer, Convolution Layer 2, pooling layer, fully connected layer, dropout layer, and output layer. This architecture is adapted from previous research [25, 29], to extract features from word or character sequences and reduce feature dimensions. The model utilizes a one-dimensional convolutional layer with TensorFlow, featuring 128 kernel filters of size 3 (Conv1D (128, 3)), as shown in Figure 3.

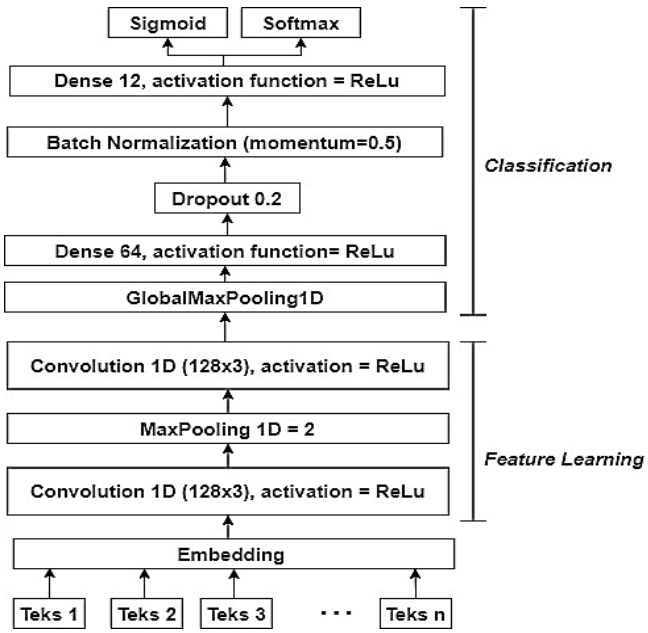

Figure 3. Architecture of the CNN model

3.4 Performance metric

The performance of the proposed method is evaluated using several metrics: precision, recall, F1-score, accuracy, and

AUC (Area Under Curve). These metrics are calculated using Eqs. (4)-(7):

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

$$Accuracy = \frac{TP+FN}{TP+FN+TN+FP} \quad (5)$$

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

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

where, *TP*, *FN*, *TN*, and *FP* correspond to true positive, false negative, true negative, and false positive, respectively. In addition, the AUC (Area Under Curve) metric is used to evaluate how well the model distinguishes between positive and negative classes (Eqs. (8) and (9)). It is calculated by plotting the Receiver Operating Characteristic (ROC) curve, which shows the relationship between True Positive Rate (TPR) and False Positive Rate (FPR):

$$TPR = \frac{TP}{(TP+FN)} \quad (8)$$

$$FPR = \frac{FP}{(FP+TN)} \quad (9)$$

These metrics provide comprehensive insights into the model's performance and its ability to accurately classify data.

4. EXPERIMENTAL RESULTS

4.1 Experimental design and hyperparameter tuning

In this study, we conduct experiments to evaluate the performance of our proposed model under different configurations. The main objective is to fine-tune the model's hyperparameters to maximize the classification accuracy for toxic speech detection. This study employs two scenarios to measure and identify the optimal model: evaluating the impact of data on overall evaluation metrics and selecting the best model parameters (Figure 4).

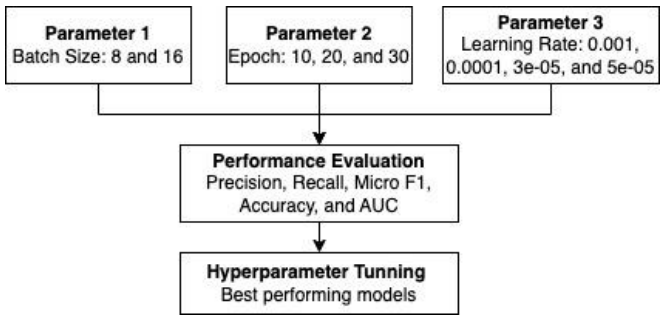

Figure 4. Training and testing scenarios to find the best model performance

The choice of hyperparameters is critical in determining the performance of deep learning models. The following hyperparameters were explored during our experiments:

- **Batch Size:** We tested batch sizes of 8 and 16. Larger batch sizes can increase computation speed but may lead to less optimal training [42]. We selected batch size 16 based on this trade-off, ensuring a balance between computational efficiency and model accuracy.
- **Learning Rate (Adam Optimizer):** We evaluated four different learning rates: 0.001, 0.0001, 3e-5, and 5e-5, which are commonly recommended for fine-tuning transformer models [20]. A larger learning rate might lead to faster convergence but risks overshooting, while smaller rates ensure stable learning but at the cost of longer training times.
- **Epochs:** We tested three epoch values: 10, 20, and 30 to observe how increasing the number of iterations impacts model performance.

In the first scenario, we implemented five data processing strategies: back translation, text expansion, word sense disambiguation (WSD), hybrid semantic enrichment (HSE), and no semantic enhancement. Each strategy achieved its best performance based on the predefined hyperparameter settings (Figure 5). A total of 24 model combinations were tested using confusion matrix evaluation.

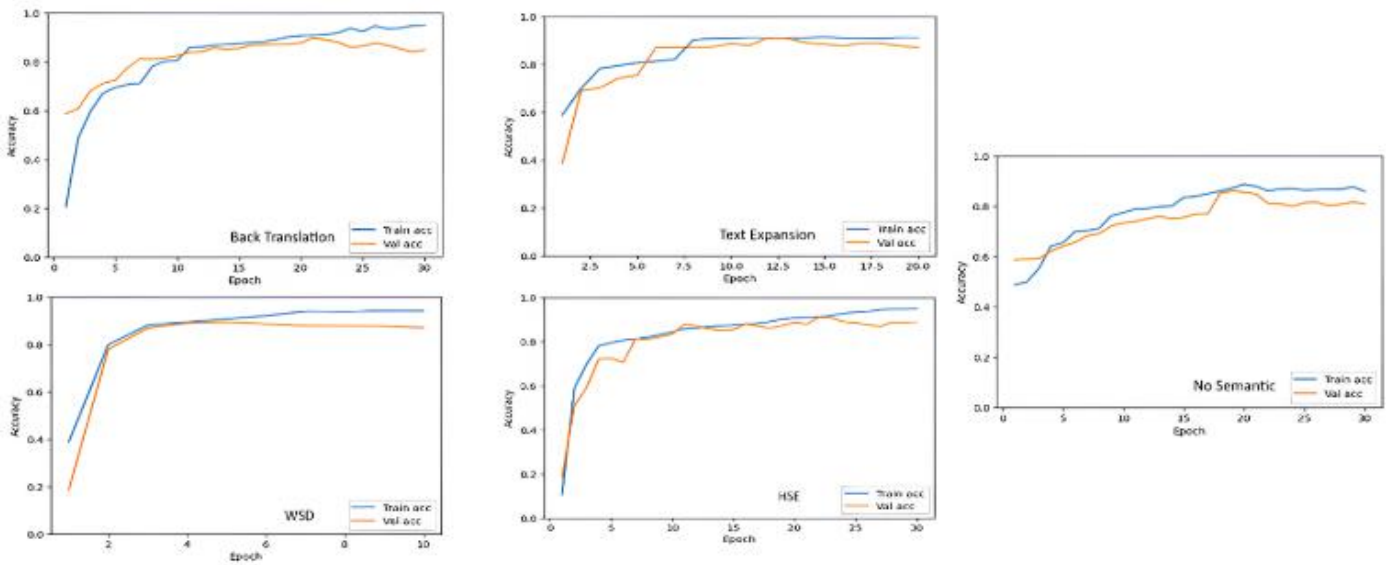

Figure 5. Analysis of the model training process with the best parameters for each strategy

The second scenario focuses on selecting the model with the best parameters. Here, the BERT embedding model is used for text vectorization due to its ability to understand and represent the Indonesian language. Each hyperparameter combination is tested using 5-fold cross-validation with a validation dataset of 1,050 tweets, resulting in 120 models per dataset. From five preprocessing strategies, a total of 600 models were evaluated. Preliminary experiments revealed that the optimal configuration (batch size 16 and learning rate 0.001) consistently yielded the highest classification accuracy and was therefore adopted in the final experimental setup.

4.2 Performance evaluation per strategy

The HSE strategy achieved the best performance with 95% training accuracy and 92% validation accuracy, due to its semantic enrichment and CNN classification approach. The WSD strategy also showed stable results (90% training, 87% validation accuracy) in just 10 epochs, highlighting the effectiveness of word sense disambiguation. Text expansion reached 90% training and 87% validation accuracy after 20 epochs. The back translation strategy performed competitively with 88% training and 85% validation accuracy. The no-semantic strategy had the lowest performance, with accuracy dropping from 89% at epoch 20 to 85% (training) and 82% (validation) at epoch 30.

Semantic enrichment in the HSE strategy significantly improved the model's understanding of textual meaning and helped manage language ambiguity through disambiguation and semantic integration. This allowed the model to handle data variation, improve contextual comprehension, and generate consistent text representations. To optimize performance and minimize loss, hyperparameter tuning was performed per model. Accuracy was used as the main metric in this multi-label classification task. For the HSE strategy, the best configuration was batch size = 16, learning rate = 0.001, and epochs = 30 in the CNN model.

4.3 Model comparison

Compared to other deep learning models such as BiGRU and BiLSTM, the CNN model consistently achieved high performance across all experimental conditions (Figure 6). With the no-semantic strategy, CNN achieved 82% accuracy, which increased to 87% with text expansion, demonstrating the positive impact of contextual enhancement.

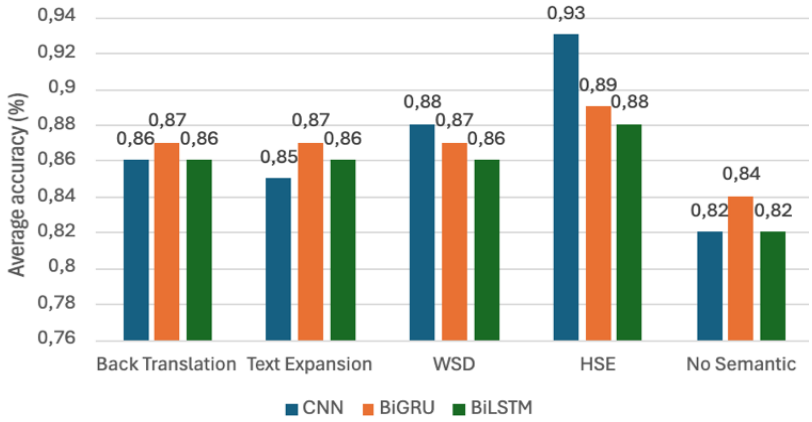

Figure 6. Comparison of classification model performance

Table 1. Proposed methodology result analysis

Strategy	k-Fold	Batch Size	Epochs	F1 Score (%)	Accuracy (%)	AUC Score (%)
Back Translation	3	16	30	64	86	84
Text Expansion	1	8	20	60	85	81
WSD	4	16	10	69	88	87
HSE	3	16	30	80	93	91
No semantic	1	16	30	60	82	79

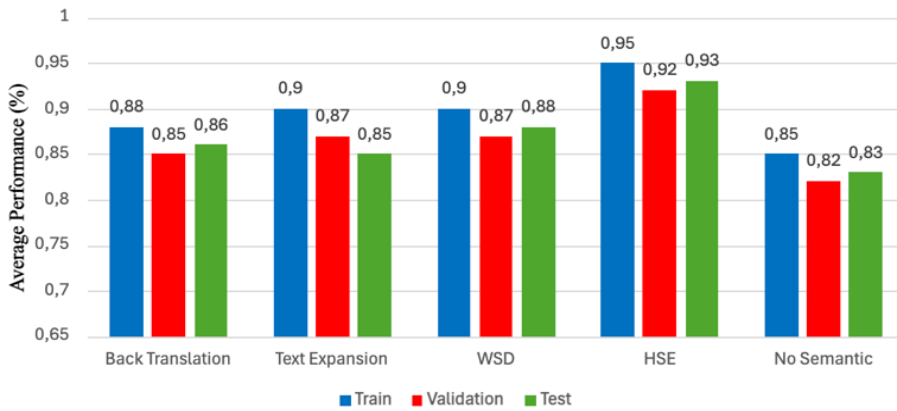

Figure 7. HSE-CNN performance analysis

Table 2. The example of tweet classification error

Sentence	Actual Label	Prediction	Prediction Error
Miss, is it better to be a prostitute or an elementary school teacher? It seems like this teacher enjoys being a slut.	toxic, abuse, individual, gender, weak	not toxic, abuse	Identification of meaning and vocabulary of the words "*prostitute*" and "*slut*," which should have been classified as toxic speech.
The weakness of Gerindra lies in this person. Prabowo's electability is declining because of Fadli Zon's mouth, like "*Durno*".	toxic, profanity, individual, other speech, weak	toxic, individual, offensive, weak	Out-of-vocabulary issue for the word "Durno," which refers to someone deceitful or dishonest (Profanity).
I'm sure they only pretend to be *212 alumni*, but in reality, they behave like infidels, aiming only to insult and divide the Muslim community through the alleged blasphemy case. They really are bastards.	toxic, abuse, group, religion, weak	toxic, profanity, individual, offensive, weak	Identification of Target and Category. The term "*212 alumni*" refers to a group, and "kavir" (out-of-vocabulary) means "*infidel*." The phrase "blasphemy case" should have been categorized appropriately.
@murniatiginting Of course, the tadpoles are going crazy for defending that frog who ruined Indonesia.	toxic, group, physical, moderate	toxic, profanity, individual offensive, weak	Identification of Target, Category, and Level. The word "*tadpole*" (an animal) should refer to a group. The word "crazy" was considered profanity. "Frog" relates more to physical appearance. "*Indonesia ruined*" should have been identified as "*Labeling*" (Moderate Level).
Looks like a scene from the *G30S-PKI* movie. It should be noted that this regime is an anti-Islam regime with no civility.	toxic, group, religion, moderate	toxic, group, religion, race, moderate	Identification of Category. The term "G30S-PKI" was identified as related to ethnic issues in history.
@investigatorri That's right, and everyone opposing the infidel Ahok will be criminalized by the police through fabricated legal cases.	toxic, individual, religion, weak	toxic, individual, offensive, strong	Identification of Category and Level. Category: the phrase "*infidel Ahok*" should be labeled under religion. Level: the word "*criminalization*" was considered a threat, increasing the severity level.

The HSE-CNN approach achieved the highest overall accuracy (93%), underscoring the contribution of semantic expansion in enhancing CNN performance. BiGRU and BiLSTM showed comparable results, with BiGRU initially outperforming CNN under non-semantic conditions. However, after semantic enrichment, BiGRU and BiLSTM reached accuracies of 89% and 88%, respectively.

4.4 HSE-CNN performance analysis

Based on Table 1 and Figure 7, the HSE-CNN model achieved the best results with 93% accuracy, 76% precision, 84% recall, 80% F1-score, and 91% AUC. The hybrid of back translation, WSD, text expansion, and semantic similarity measurement improved contextual understanding and accuracy.

4.5 Label-wise performance analysis

Semantic enrichment contributes positively by improving contextual understanding and enabling accurate predictions. However, not all labels achieved equally high accuracy due to semantic overlap, limited data, or ambiguity. Misclassifications often arise in cases where toxic speech is context-dependent, ambiguous, or closely related to non-toxic expressions, leading to false positives or false negatives in prediction (see Table 2). For example, labels such as *physical* (84%) and *offensive* (88%) in the HSE model demonstrate higher performance.

- High-performing labels (e.g., toxic, abuse, individual, group, level): benefited from consistent and abundant data.
- Moderate performance labels (e.g., religion, race, offensive): faced prediction inconsistency due to overlapping semantics.
- Low performance labels (e.g., physical, gender): hindered by sparse and ambiguous data.

4.6 Comparative analysis with existing methods

Compared to previous methods, HSE-CNN demonstrates superior accuracy and robustness, particularly in multi-label toxic speech classification. Despite the promising results, label imbalance in the dataset introduced inconsistencies in precision, recall, F1-score, and accuracy, particularly affecting the classification of minority categories. This issue often led to biased learning behavior, where the model favored more frequent labels while underperforming on rare ones, such as label Physical, Gender, and Other.

To mitigate this, we applied a class weighting technique during training, assigning greater loss penalties to underrepresented labels. This approach aimed to improve the model's sensitivity toward minority classes by adjusting the learning process to counteract label frequency imbalance. However, due to technical constraints in TensorFlow's multi-label loss handling, the weighting could only be partially implemented. Still, its effect on the few classes tested was measurable and positive.

Table 3 provides an example of the impact of class weighting on the model's performance with respect to minority classes. In the table, we show the F1-scores before and after class weighting for three underrepresented labels: Physical, Gender, and Other. As can be seen, the F1-score of these labels increased noticeably after applying class weighting, with improvements of +0.04 for Physical, +0.03 for Gender, and +0.04 for Other, indicating a positive effect of the weighting technique on these minority categories.

Table 3. Example of class weighting's impact on minority-class F1-scores

Label	F1 (No Weighting)	F1 (Weighting)	ΔF1
Physical	0.74	0.78	+0.04
Gender	0.73	0.76	+0.03
Other	0.76	0.80	+0.04

Table 3 exemplifies how class weighting can enhance the performance of models in the presence of label imbalance, though it is important to note that this is only one example. Further refinements, such as the application of focal loss, oversampling techniques (e.g., SMOTE), or cost-sensitive learning, are recommended for future research to better address the imbalance, especially in complex multi-label tasks.

Real-time testing also revealed that short, ambiguous texts and overlapping toxic speech categories posed significant challenges. In these cases, classification performance could degrade due to label confusion, demonstrating the need for additional work in semantic disambiguation and augmentation techniques to further improve model accuracy.

5. DISCUSSION

The development of a toxic speech classification model has been successfully carried out using a hybrid semantic enrichment strategy combined with a Convolutional Neural Network (HSE-CNN). The model leverages the lexical database WordNet for semantic feature engineering, enabling the grouping of words into synsets and capturing lexical relationships such as synonyms, antonyms, hyponyms, and meronyms. This semantic enhancement has proven essential for improving classification performance, especially in handling the contextual complexity of Indonesian multi-label toxic speech data. The proposed model achieved a notable performance of 93% accuracy using the HSE-CNN approach.

Before modeling, semantic analysis of unstructured text data from X (formerly Twitter) was conducted, emphasizing preprocessing steps such as case folding, normalization, and back-translation. These steps ensure the structural integrity of input sentences and help the model better interpret context and meaning. Additionally, contextual embeddings such as BERT were used to represent semantic information at the sentence level. Label definitions were adapted from prior studies on toxic speech in Indonesian language contexts [32], simplifying from 12 to 9 labels for better label balance and clearer categorization.

Hyperparameter tuning was performed using K-Fold Cross-Validation, resulting in optimal settings: a learning rate of 0.001, batch size of 16, and 30 epochs. The HSE-CNN achieved 76% precision, 84% recall, 80% F1-score, 93% accuracy, and 91% AUC, surpassing other semantic enrichment baselines such as back translation (F1-score 64%, accuracy 86%), semantic text expansion (F1-score 60%, accuracy 85%), and expansion with disambiguation (F1-score 69%, accuracy 88%).

Table 4. Comparative analysis of toxic detection model with state of art

Method	Number of Labels	Avg.Acc
BiGRU + IndoBERT [43]	2	83,8%
SVM [32]	9	68,3%
BiLSTM + BERT [29]	12	64,8%
RFDT+CC [29]	12	76,1%
CNN+DistilBERT [44]	12	61,3%
SVM+CC [44]	12	74,8%
Word2Vec+BiLSTM [45]	12	80,25%
IndoBERT [46]	12	88,25%
HSE+CNN (**Proposed)**	**12**	**93%**

To further assess model performance, additional experiments were conducted with BiGRU and BiLSTM models using the same settings. Results showed CNN consistently outperformed both, with a 4% and 5% accuracy margin, respectively, reaffirming CNN's strengths in processing parallel inputs and extracting local features.

The comparative performance of HSE-CNN with other state-of-the-art models is shown in Table 4.

Previous models such as BiGRU + IndoBERT [43] and SVM [32] were either limited to binary classification or showed lower accuracy for multi-label classification. While IndoBERT alone [46] achieved strong results, the integration of semantic enrichment in HSE-CNN provides a substantial improvement. Studies like [29, 44] also employed translation and classifier chains, with limited effectiveness for multi-label scenarios, confirming the advantage of the proposed HSE-CNN approach.

Although the proposed HSE-CNN approach demonstrates high performance under experimental conditions, several practical limitations must be acknowledged for real-world deployment, particularly in social media environments:

- Computational Efficiency: The semantic enrichment steps (including back translation, text expansion, and disambiguation), introduce significant computational overhead. These steps are not optimized for real-time processing and may hinder deployment in large-scale social media monitoring systems.
- Real-Time Application: While CNN itself is efficient during inference, the entire HSE-CNN pipeline is not currently suitable for real-time toxicity detection. For operational deployment, it may require optimization strategies such as model quantization, knowledge distillation, or switching to lightweight transformer variants (e.g., DistilBERT, TinyBERT).
- Scalability: Real-world applications on platforms like X (Twitter) require attention to API latency, memory efficiency, and throughput. The current approach, although effective, needs streamlining of preprocessing and enrichment stages to meet performance demands at scale.
- Language Evolution: Toxic language in social media is constantly evolving. The static training of the current model may result in decreased performance over time unless it is periodically retrained or integrated into continual learning frameworks.

Future work will address these challenges by focusing on the optimization of semantic enrichment modules, real-time inference efficiency, and adaptive learning techniques. These enhancements are necessary to bridge the gap between research and deployment in live moderation systems on social media platforms.

6. CONCLUSIONS AND FUTURE WORK

This study successfully developed a novel approach, Hybrid of Semantic Enrichment and Convolutional Neural Network (HSE-CNN), which enhances the performance of toxic speech classification on multi-label datasets. The HSE-CNN approach enriches the semantic understanding of text, enabling the model to capture nuances and context more effectively. By integrating semantic information with CNN models, this approach achieves impressive performance metrics: precision of 76%, recall of 84%, F1-Score of 80%, accuracy of 93%, and AUC of 91%.

HSE-CNN surpasses other approaches such as back translation, text expansion, and text disambiguation. Moreover, it outperforms deep learning models like BiGRU and BiLSTM. Previous studies have demonstrated that the HSE approach enhances classification performance, particularly on datasets with multiple labels.

While the HSE-CNN approach demonstrates success in improving toxic classification performance, this study identifies several limitations that warrant attention. Firstly, data label imbalance may affect model performance outcomes. Therefore, it is advisable to explore advanced techniques for handling label imbalance in multi-label datasets, such as oversampling or undersampling strategies. Secondly, utilizing HSE-CNN requires significant computational resources and time. Hence, optimizing computational processes through model optimization techniques or selecting more efficient resources is crucial.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Bina Darma University, the funding institution that made the publication of this research possible.

REFERENCES

[1] Modha, S., Majumder, P., Mandl, T. (2022). An empirical evaluation of text representation schemes to filter the social media stream. Journal of Experimental & Theoretical Artificial Intelligence, 34(3): 499-525. https://doi.org/10.1080/0952813X.2021.1907792

[2] Jonathan, V.W., Setiawan, E.B. (2023). Feature expansion using GloVe for hate speech detection using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) method in Twitter. In 2023 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, pp. 197-202. https://doi.org/10.1109/ICoDSA58501.2023.10277204

[3] Ghozali, I., Sungkono, K.R., Sarno, R., Abdullah, R. (2023). Synonym based feature expansion for Indonesian hate speech detection. International Journal of Electrical and Computer Engineering (IJECE), 13(1): 1105. https://doi.org/10.11591/ijece.v13i1.pp1105-1112

[4] Nabiilah, G.Z., Alam, I.N., Purwanto, E.S., Hidayat, M.F. (2024). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. International Journal of Electrical & Computer Engineering (2088-8708), 14(1): 1071-1078. https://doi.org/10.11591/ijece.v14i1.pp1071-1078

[5] Ibrohim, M.O., Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. Heliyon, 9(8): e18647. https://doi.org/10.1016/j.heliyon.2023.e18647

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

[7] Wang, H., Tian, K., Wu, Z., Wang, L. (2021). A short text classification method based on convolutional neural network and semantic extension. International Journal of Computational Intelligence Systems, 14(1): 367-375. https://doi.org/10.2991/ijcis.d.201207.001

[8] 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

[9] Niam, I.M.A., Irawan, B., Setianingsih, C., Putra, B.P. (2018). Hate speech detection using latent semantic analysis (LSA) method based on image. In 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Bandung, Indonesia, pp. 166-171. https://doi.org/10.1109/ICCEREC.2018.8712111

[10] Pradhan, A., Senapati, M.R., Sahu, P.K. (2022). Improving sentiment analysis with learning concepts from concept, patterns lexicons and negations. Ain Shams Engineering Journal, 13(2): 101559. https://doi.org/10.1016/j.asej.2021.08.004

[11] Datchanamoorthy, K. (2023). Text mining: Clustering using bert and probabilistic topic modeling. Social Informatics Journal, 2(2): 1-13. https://doi.org/10.58898/sij.v2i2.01-13

[12] Wulandari, W., Makmur, H., Surianto, D.F., Risal, A.A.N., Budiarti, N.A.E., Zain, S.G., Wahid, A. (2025). Semantic feature engineering with LSA-SVM for cyberbullying comment classification on instagram. Informatica, 49(15): 165-178. https://doi.org/10.31449/inf.v49i15.6992

[13] Zouidine, M., Khalil, M. (2025). Selective reading for Arabic sentiment analysis. IEEE Access, 13: 59157-59169. https://doi.org/10.1109/ACCESS.2025.3556976

[14] 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.

[15] Beddiar, D.R., Jahan, M.S., Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media, 24: 100153. https://doi.org/10.1016/j.osnem.2021.100153

[16] Maryamah, M., Arifin, A.Z., Sarno, R. (2019). Query expansion based on Wikipedia word embedding and BabelNet method for searching Arabic documents. International Journal of Intelligent Engineering & Systems, 12(5): 202-213. https://doi.org/10.22266/ijies2019.1031.20

[17] Sharma, D.K., Pamula, R., Chauhan, D.S. (2021). Semantic approaches for query expansion. Evolutionary Intelligence, 14(2): 1101-1116. https://doi.org/10.1007/s12065-020-00554-x

[18] Mikolov, T., Yih, W.T., Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746-751.

[19] Anjum, Katarya, R. (2024). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. International Journal of Information Security, 23(1): 577-608. https://doi.org/10.1007/s10207-023-00755-2

[20] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019).

Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

[21] El-Razzaz, M., Fakhr, M.W., Maghraby, F.A. (2021). Arabic gloss WSD using BERT. Applied Sciences, 11(6): 2567. https://doi.org/10.3390/app11062567

[22] Elekes, Á., Schäler, M., Böhm, K. (2017). On the various semantics of similarity in word embedding models. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada, pp. 1-10. https://doi.org/10.1109/JCDL.2017.7991568

[23] Zingla, M.A., Chiraz, L., Slimani, Y. (2016). Short query expansion for microblog retrieval. Procedia Computer Science, 96: 225-234. https://doi.org/10.1016/j.procs.2016.08.135

[24] Zhou, Y., Xu, J., Cao, J., Xu, B., Li, C. (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

[25] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, pp. 1746-1751. https://doi.org/10.3115/v1/D14-1181

[26] Maslej-Krešňáková, V., Sarnovský, M., Butka, P., Machová, K. (2020). Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Applied Sciences, 10(23): 8631. https://doi.org/10.3390/app10238631

[27] Bagate, B.A., Joshi, A.S., Kadam, A., Choubey, C.K., Sable, N., Kumar, A., Dogra, N., Nandan, D. (2025). Sarcasm detection an explainable AI approach for reddit political text. Mathematical Modelling of Engineering Problems, 12(1): 219-226. https://doi.org/10.18280/mmep.120123

[28] Alfina, I., Mulia, R., Fanany, M.I., Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. In 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, Indonesia, pp. 233-238. https://doi.org/10.1109/ICACSIS.2017.8355039

[29] Hendrawan, R., 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), Bandung, Indonesia, pp. 1-7. https://doi.org/10.1109/ICoDSA50139.2020.9212962

[30] Rahmawati, D., Khodra, M.L. (2016). Word2vec semantic representation in multilabel classification for Indonesian news article. In 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA), Penang, Malaysia, pp. 1-6. https://doi.org/10.1109/ICAICTA.2016.7803115

[31] 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, Florence, Italy, pp. 46-57. https://doi.org/10.18653/v1/w19-3506

[32] 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), Semarang, Indonesia, pp. 1-5. https://doi.org/10.1109/ICITACEE.2019.8904425

[33] Buey, M.G., Garrido, Á.L., Ilarri, S. (2014). An approach for automatic query expansion based on NLP and semantics. In Database and Expert Systems Applications: 25th International Conference, DEXA 2014, Munich, Germany, pp. 349-356. https://doi.org/10.1007/978-3-319-10085-2_32

[34] Ansari, G., Kaur, P., Saxena, C. (2024). Data augmentation for improving explainability of hate speech detection. Arabian Journal for Science and Engineering, 49(3): 3609-3621. https://doi.org/10.1007/s13369-023-08100-4

[35] Siino, M., Lomonaco, F., Rosso, P. (2024). Backtranslate what you are saying and I will tell who you are. Expert Systems, 41(8): e13568. https://doi.org/10.1111/exsy.13568

[36] Carpineto, C., Romano, G. (2012). A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR), 44(1): 1-50. https://doi.org/10.1145/2071389.2071390

[37] Voorhees, E.M. (1994). Query expansion using lexical-semantic relations. In SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Organised by Dublin City University, pp. 61-69. https://doi.org/10.1007/978-1-4471-2099-5_7

[38] Azad, H.K., Deepak, A. (2019). A new approach for query expansion using Wikipedia and WordNet. Information Sciences, 492: 147-163. https://doi.org/10.1016/j.ins.2019.04.019

[39] Shruthi, P., KM, A.K. (2020). Novel approach for generating hybrid features set to effectively identify hate speech. Inteligencia Artificial, 23(66): 97-111. https://doi.org/10.4114/intartif.vol23iss66pp97-111

[40] Alsafari, S., Sadaoui, S., Mouhoub, M. (2020). Deep learning ensembles for hate speech detection. In 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, pp. 526-531. https://doi.org/10.1109/ICTAI50040.2020.00087

[41] Chen, J., Hu, Y., Liu, J., Xiao, Y., Jiang, H. (2019). Deep short text classification with knowledge powered attention. Proceedings of the AAAI conference on artificial intelligence, 33(1): 6252-6259. https://doi.org/10.1609/aaai.v33i01.33016252

[42] 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. JIEET (Journal of Information Engineering and Educational Technology), 5(2): 44-48. https://doi.org/10.26740/jieet.v5n2.p44-48

[43] 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), Bangsaen, Chonburi, Thailand, pp. 186-190.

https://doi.org/10.1109/KST51265.2021.9415760
[44] Hana, K.M., 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), Bandung, Indonesia, pp. 1-7. https://doi.org/10.1109/ICoDSA50139.2020.9212992
[45] Zahra, E.A.A., Sibaroni, Y., Prasetyowati, S.S. (2023). Classification of multi-label of hate speech on twitter Indonesia using LSTM and BiLSTM method. JINAV: Journal of Information and Visualization, 4(2): 170-178. https://doi.org/10.35877/454RI.jinav1864
[46] Darmawan, M.A., Boentoro, N.W., Surya, K.C., Sutoyo, R. (2023). Experiments on indobert implementation for detecting multi-label hate speech with data resampling through synonym replacement method. In 2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Kuala Lumpur, Malaysia, pp. 1-7. https://doi.org/10.1109/ICRAIE59459.2023.10468099