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

Ingénierie des Systèmes d'Information
Vol. 28, No. 3, June, 2023, pp. 545-556
Journal homepage: http://iieta.org/journals/isi

Advancements in Semantic Expansion Techniques for Short Text Classification and 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, Universitas Bina Darma, 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/isi.280302

Received: 20 May 2023
Accepted: 10 June 2023

Keywords:
Indonesia hate speech detection, semantic expansion, semantic similarity, short text expansion, text classification

ABSTRACT

Traditional text classification methodologies, which primarily rely on document context and word frequency, often fall short in handling the linguistic complexities of the Indonesian language, such as colloquialism and informal language usage. This study presents a comprehensive semantic expansion-based framework to address these challenges in detecting hate speech within Indonesian social media commentary. Our framework leverages trusted knowledge bases, WordNet and Kateglo, to alleviate ambiguity in short texts. The BERT word insertion model is employed for semantic similarity calculation, followed by the application of a CNN deep learning model for hate speech classification. This approach effectively enhances semantic understanding and accurately classifies hate speech. The study also highlights future trajectories in semantic expansion for short text classification, encouraging further research to implement the proposed framework as an automated detection system.

1. INTRODUCTION

1.1 Characteristics of hate speech and associated challenges

The proliferation of digital information has precipitated an uncontrolled surge in the dissemination of ideas, particularly on social media platforms. The diversity of languages from which social media content originates presents formidable challenges in preprocessing and detection. Furthermore, social media's facilitation of rapid, concise communication has had a significant impact on freedom of expression [1]. Nevertheless, the brevity of social media messages often results in incomplete information, the use of ambiguous abbreviations, and the employment of complex language [2]. These factors adversely affect analysis and detection processes, introducing complications such as word correlation, noise, and ambiguity, which impede the efficacy of text classification algorithms. Moreover, certain factions exploit social media to provoke, express biased opinions, disseminate slander, and incite hatred towards specific individuals or groups [3]. Hate speech typically targets various categories including religion, race, physical appearance, and gender, among others (Figure 1) [4].

For instance, consider the following Indonesian hate speech example: "lu semua dr etnis mata sipit harusnya jd kacung atau babu saja". This sentence is characterized by grammatical inaccuracies, ambiguous phrases, and noise. For improved clarity and contextual understanding, consider the revised sentence: "Kamu semua dari etnis mata sipit seharusnya menjadi kacung atau pembantu saja (All of you from the narrow-eyed ethnicity can only be lackeys or helpers)".

According to the hate speech guidebook released by the National Commission on Law and Human Rights (Komnas HAM) in Indonesia [5], hate speech is typically directed towards specific entities and can be classified into two categories: individual or group targets. Hate speech is further delineated into distinct groups, excluding categories such as profanity or slander [4]. Furthermore, hate speech is classified into varying intensity levels: weak, moderate, and strong. However, to conduct a comprehensive classification encompassing multiple groups, a clear contextual understanding of each sentence is indispensable.

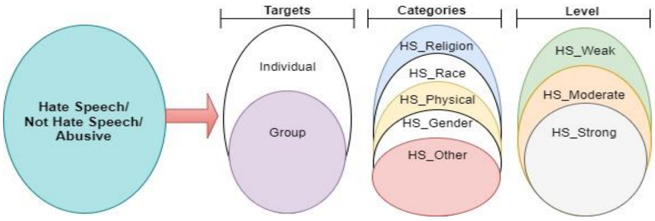


Figure 1. Classification of hate speech in Indonesia

1.2 Challenges in short text classification and semantic solutions

Text classification, particularly concerning short texts, poses considerable challenges in the selection of suitable methods and models. Given their complexity and conciseness, short texts, such as social media tweets, comments, or messages, often result in suboptimal model performance and lower classification accuracy [6]. A comprehensive understanding of these texts is paramount for interpreting user intent [7]. However, the inherent limitations of short text classification, including insufficient contextual information, result in issues such as data dispersion and high degrees of ambiguity [8]. Traditional techniques such as bag-of-words (BoW) fail to address these challenges adequately due to their inability to account for word order and semantic relationships [9].

To overcome these obstacles, researchers have proposed innovative frameworks such as the one developed by Collobert et al. [10], which employs latent Dirichlet allocation (LDA) in conjunction with Wikipedia to tackle the unique challenges posed by short texts. Other scholars have adopted a semantic approach to mitigate these challenges, focusing on understanding word meanings and uncovering hidden semantic relationships [11-14]. By capturing these elements, the semantic approach aids in the accurate categorization of short texts via information retrieval [13].

To augment information retrieval, semantic expansion techniques have been proposed to bridge knowledge gaps [15, 16]. For instance, the study made by Diaz et al. [12] implemented a query expansion (QE) technique that incorporated synonyms, hyponyms, and hypernyms of words in a user's query through lexical analysis. This method has demonstrated efficacy in reducing term mismatches and has been widely endorsed [6]. The issue of term mismatch often arises when users struggle to identify suitable search terms, and automatic query expansion (AQE) can alleviate this problem.

Various AQE techniques have been classified into local and global approaches by researchers, and commonly rely on ontologies or other semantic-based sources such as WordNet or thesauri for query expansion. Additionally, several semantic enhancement methods have been developed, broadly categorized into ontology-based, linguistic-based, and hybrid approaches [12, 17]. Knowledge sources such as WordNet, Wikipedia, and Wiktionary are frequently utilized in these approaches [12, 17]. In the research conducted by Devi and Gandhi [18], a combined corpus-based technique and a relevance feedback mechanism were applied for AQE, calculating semantic similarity using word embeddings and comparing knowledge differences between user queries and feature extraction results. In another study, Azad and Deepak [19] employed natural language processing (NLP) and multilingual semantics to restructure user requests, thereby enhancing information extraction from data repositories.

1.3 Motivation and research objectives

The issues of polysemy, data dispersion, and ambiguity in text classification have been extensively examined, with a range of methodologies developed to integrate semantic relationships into these processes. However, the application of semantic expansion techniques to short text classification continues to present significant challenges, warranting further research. As such, this study is instrumental in investigating the impact of semantic expansion on the classification of short texts, particularly in the context of hate speech.

Short texts, due to their brevity, pose challenges in rapid comprehension compared to lengthier documents [6]. The application of semantic expansion techniques often exacerbates this issue, leading to a loss of context and increased ambiguity. Nevertheless, these techniques can play a pivotal role in overcoming the inherent contextual limitations of short texts, through the utilization of knowledge bases such as WordNet and other thesauri. Therefore, it is of paramount importance to assess the efficacy of these techniques in detecting hate speech on Indonesian social media platforms.

The objectives of this study are twofold: (1) to explore the application of semantic expansion techniques to short text classification, and (2) to propose a semantic-based framework for hate speech detection in Indonesian social media comments.

In light of the above, this paper makes the following significant contributions:

- The necessity of semantic expansion in addressing hate speech, especially in the context of short texts, is underscored. Furthermore, a comprehensive examination of existing resolution techniques is provided.
- A novel framework that integrates various knowledge-based methods and word embeddings within a semantic approach is proposed. The framework is designed to enhance sentence context and reduce word ambiguity.

In summary, the proposed framework employs semantic expansion to facilitate short text classification and hate speech detection, paving the way for the development of an automated monitoring system. With further advancements, this approach could be extended to other relevant domains, such as cyberbullying detection and abusive content moderation.

2. TEXT CLASSIFICATION BASED ON SEMANTIC EXPANSION APPROACH

Semantic expansion strategies pivot on the identification of words or phrases pertinent to the user's objectives. These relevant terms are incorporated into the initial search results via the mechanism of query expansion (QE) [20]. The primary aim of QE is to mitigate the inherent ambiguity in natural language processing and to furnish a more comprehensive insight by enriching the original query with pertinent terms. Traditional QE methodologies, encompassing global and local analysis, lean on the corpus to broaden search queries [21]. Candidate terms are gleaned through statistical scrutiny of the corpus content. However, as posited by Bhogal et al. [22], corpus-based statistical approaches yield effective results only when the corpus in question is sufficiently voluminous and pertains to the domain of the search query. Contrastingly, the semantic expansion methodology is not bound by such constraints, given its reliance on knowledge structures independent of the corpus, such as a thesaurus or an ontology.

Approaches based on semantic text expansion can be bifurcated into two primary categories: global analysis and local analysis. A comprehensive summary of all literature studies pertinent to the expansion approach has been compiled in Table 1. Additionally, an exhaustive review has been provided in Table 2, outlining the advantages and disadvantages of each approach. The ensuing discussion on global and local analysis will be further split into four and two categories, respectively, as depicted in Figure 2. However, the immediate discussion will be confined to the two categories under global analysis, specifically knowledge-based and word embedding.

Table 1. Summary of literature studies related to the semantic expansion approach

Semantic expansion approach	Ref.	Knowledge structure	Approach	Dataset	Performance metrics	Result
Knowledge -based	[23]	Wikipedia	Improving weak Ad-hoc queries using PRF	TREC, BigNews	MAP	MAP = 0.2094
	[24]	WordNet	Weighting terms	News sources	MAP	MAP= 0.206, Recall= 0.895
	[25]	WordNet	Expands the query by incorporating synonyms and extracting linguistic phrases from source code identifiers.	45KLOC JavaScript/ ECMAScript	Precision (P), Recall (R)	P=0.87, R=0.86
	[26]	ConceptNet	Builds a fuzzy ontology by integrating text mining techniques, external ontologies, and the global ontology ConceptNet.	NewsGroup20	MAP, MRR, Precision	MAP: @Google=0.93, @Bing=0.89
Word embedding-based	[8]	Glove	The logistic bilinear regression model utilizes both global matrix factorization (LSA) and local context window methods (Skip-Gram).	CoNLL-2003 dataset	Accuracy	0.829
	[12]	Word2vec & Glove	The relatedness of terms in the context of query extension for ad hoc information retrieval.	TREC ad hoc	Qualitative: Spearman's Correlation	0.563
	[14]	Word2vec	Global and Local Word Embedding-Based Topic Model (GLTM) using the Continuous Skip-gram model with Negative Sampling (SGNS).	Web and microblogs	Accuracy	0.85
	[27]	BERT	Capture the meaning of each word in the context of a sentence or document.	Baidu Chinese	Accuracy	F1 score of = 0.8434
	[16]	BERT	Contextual word insertion	TREC	MAP	0.2253
	[28]	BERT	BERT is used to calculate semantic similarity scores and term scores at the sentence level, and then determine the weight of terms and sentences using the Rocchio model.	TREC ad hoc	MAP, P@10, MRR dan NDCG	MAP= @TREC_ GOV2 0.7613
Hybrid-based	[18]	WordNet & ontologi	WordNet formulates user queries, and ontologies improve query expansion results.	News sources	Accuracy	0.871
	[29]	ConceptNet & WordNet	The query is expanded by selecting candidate terms.	Document	MAP	0.2297
	[30]	Word Embedding & BableNet	Candidates are developed by searching for synonyms, measuring similarity with WordNet and word embedding, as well as BableNet embedding.	News sources	Accuracy	0.890

Table 2. The pros and cons of semantic expansion approaches

Semantic expansion approach	Pros	Cons
Knowledge-based approach	Producing accurate and memorable results with high precision and good recall.	Performance depends on the quality of ontology knowledge.
Word embedding-based approach	Capable of considering the semantic relationships between words. Can capture the context-dependent meaning of a word.	Selecting relevant words for query expansion is not always straightforward.
Hybrid-based approach	Integrates the benefits of multiple query expansion methods.	Necessitates greater computational resources.

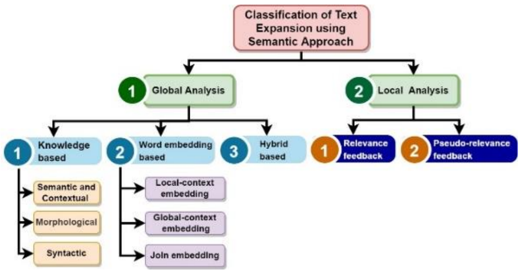


Figure 2. Classification of semantic approaches

2.1 Knowledge-based approach

In the knowledge-based approach, an analysis of various linguistic features, including the morphological, semantic, and syntactic relations of query terms, is conducted. This approach entails the replacement of query terms with words holding similar contextual meanings, accomplished through the utilization of diverse knowledge sources such as thesauri, dictionaries, ontologies, and cloud-based Linked Open Data (LOD) resources like WordNet, Wiktionary, Wikipedia, or

ConceptNet [31]. Hence, the expansion of knowledge-based queries is categorized into three sections, grounded in previous studies: semantics and context, morphology, and syntax.

2.1.1 Semantics and context

Semantic and contextual approaches involve utilizing various knowledge sources such as ontologies, cloud LOD, dictionaries, and thesauri. WordNet, for instance, is a thesaurus that defines semantic relationships between words, encompassing synonyms, hyponyms, and meronyms. Due to its effectiveness in addressing ambiguity, researchers have adopted this approach in their studies. WordNet is employed to discover synonyms [27], while Harman [32] utilizes Wikipedia and WordNet as data sources for query expansion, considering term weights. In another article [33], POS tags are used to query words and extract their synonyms from the WordNet lexicon. Furthermore, recent studies have incorporated ConceptNet as a data source for QE. For example, Storey et al. [24] employ ConceptNet's Global Ontology to tackle ambiguity. Additionally, Wikipedia has gained widespread usage as a data source, as demonstrated in the study [34], where it is utilized to acquire additional terms for pseudo-relevance feedback.

To exemplify, we can leverage WordNet to expand words and find synonyms. WordNet encompasses synsets, enabling us to identify words and definitions. For instance, given the input “kamu sangat bodoh (you are so stupid)” as a short text, we can employ WordNet to find synonyms for "kamu" and relevant verbs or adjectives for "bodoh" in different contexts. However, it's important to note that WordNet does not specifically recognize complete phrases or sentences; rather, it focuses on individual words. Therefore, in the context of the sentence "kamu sangat bodoh", we can combine WordNet synonyms for "kamu" with existing synonyms for "bodoh". For instance, synonyms for "kamu" may include "anda" or "dirimu" while synonyms for "bodoh" may consist of "goblok" or "tolol". Consequently, the result of semantic expansion could be variations like "anda sangat goblok" or "dirimu sangat tolol" conveying the same meaning.

2.1.2 Morphological

Morphological expansion is a technique that involves utilizing various forms of keywords, including root words, parts of speech, adjectives, and progressive tenses. For instance, word stemming, which is one of the earliest query expansion (QE) approaches, remains influential in reducing inflectional forms to their root words [35]. Additionally, a study by the study [36] has demonstrated that incorporating morphological variations in the debriefing system enhances memory. Furthermore, in another study [23], expanding keywords with morphological variants extracted from documents significantly improves information retrieval performance. This technique has led to the development and successful application of stemmers such as Paice/Husk, Krovetz, and Porter in various QE strategies, aiming to enhance accuracy and precision in information retrieval [37-40].

To illustrate the concept of stemming, let's consider the word “bakar (burn)” and its various forms, including [‘membakar’, ‘terbakar’, ‘dibakar’, ‘pembakaran’]. Stemming is performed by reducing these unique words to their root form, thereby improving performance by removing prefixes and suffixes. Similarly, when applying morphological expansion to the phrase "kamu sangat bodoh (you are so stupid)", we can generate different tense variations suitable for various contexts. An example of morphological expansion for this phrase could be: (1) “Anda sangat bodohkan” (where the morphology of the word "kamu" changes to "anda" and " bodoh" changes to " bodohkan" takes the verb form with the addition of a suffix) and (2) Dirimu sangat bodoh (where the morphology of the word "kamu" changes to "dirimu" without any change to "bodoh").

2.1.3 Syntactic

Syntax utilizes enhanced relational features of query terms to expand the original query through statistical approaches. It uses syntax and lexicons to obtain noun phrases, terminologies, and entities from documents, thereby enhancing representation with linguistic units from the knowledge base [41]. For example, in the study [15], the focus is on the use of multi-word phrases for text representation through noun phrase syntax structure, proposing two strategies for representing general concepts and subtopic representations to represent documents. Additionally, in the study [42], NLP techniques such as named entity recognition (NER) and syntactic parsing have been employed to build semantic and syntactic structures of the corpus.

As an illustration, syntactic expansion for the phrase "kamu sangat bodoh (you are so stupid)" involves changes in the sentence's structure or syntax. The word "Kamu" can become [‘Anda’, ‘Kalian’, ‘Dirimu’], the word "Sangat" remains unchanged, and the word "Bodoh" can consider changes in its position. Therefore, examples of syntactic expansions for the phrase "kamu sangat bodoh" could be: (1) “Bodoh sekali kamu” (changing the position of the word "bodoh" in front of the word "kamu"), (2) “Kamu benar-benar bodoh” (using the construction "benar-benar" as an intensifier), or (3) “Kamu sangat bodoh, ya” (adding the particle "ya" as a sentence follower to indicate conviction).

2.2 Word embedding-based approach

In the knowledge-based approach, an analysis of various linguistic features, including the morphological, semantic, and syntactic relations of query terms, is conducted. This approach entails the replacement of query terms with words holding similar contextual meanings, accomplished through the utilization of diverse knowledge sources such as thesauri, dictionaries, ontologies, and cloud-based Linked Open Data (LOD) resources like WordNet, Wiktionary, Wikipedia, or ConceptNet [31]. Hence, the expansion of knowledge-based queries is categorized into three sections, grounded in previous studies: semantics and context, morphology, and syntax.

2.2.1 Local embedding

The semantics and context approach involves the use of a variety of knowledge sources such as ontologies, cloud LOD, dictionaries, and thesauri. For instance, WordNet is a thesaurus that delineates semantic relationships between words, covering synonyms, hyponyms, and meronyms. Given its effectiveness in resolving ambiguity, this approach has been adopted in various research studies. WordNet is used to identify synonyms [27], whereas Harman [32] employs Wikipedia and WordNet as data sources for query expansion, considering term weights. In another study, Salton and Buckley [33] leverage POS tags to query words and extract their synonyms from the WordNet lexicon. Moreover, recent studies have incorporated ConceptNet as a data source for QE.

For example, Storey [24] utilizes ConceptNet's Global Ontology to address ambiguity. Wikipedia has also seen widespread usage as a data source [34], where it is employed to gain additional terms for pseudo-relevance feedback.

To illustrate, WordNet can be employed to expand words and discover synonyms. WordNet comprises synsets, facilitating the identification of words and definitions. Given the input “kamu sangat bodoh (you are so stupid)” as short text, WordNet can be used to identify synonyms for "kamu" and relevant verbs or adjectives for "bodoh" in different contexts. However, it must be noted that WordNet does not specifically recognize complete phrases or sentences, focusing instead on individual words. Thus, in the context of the sentence "kamu sangat bodoh", WordNet synonyms for "kamu" can be combined with existing synonyms for "bodoh". For example, synonyms for "kamu" might include "anda" or "dirimu", while synonyms for "bodoh" could be "goblok" or "tolol". Consequently, the outcome of semantic expansion could entail variations such as "anda sangat goblok" or "dirimu sangat tolol", imparting the same meaning.

2.2.2 Global embedding

Recently, an implicit incorporation of a global context in the network structure has been observed in studies investigating word embedding. For instance, local embedding has been employed [12] to capture topics, yielding superior outcomes compared to global embedding. Moreover, Doc2Vec, a technique that leverages word pair models to predict words within documents, was introduced [13] to facilitate the learning of sentence and document representations.

2.2.3 Combined or mixed embedding

Efforts have been made to integrate both local and global contexts into word embedding. A Topic Model Based on Global and Local Word Embedding (GLTM) for short texts is introduced [14]. This model utilizes both global and local word embedding to filter semantic connections between words. For example, in study, a query expansion method based on contextual word insertion using BERT word embedding is proposed [16]. The intention is to enhance document ranking performance by selecting the most suitable candidate term from a set of terms. Interestingly, it was discovered that the BERT model outperformed Word2Vec. In addition, a probabilistic framework combining BERT's sentence-level semantics with a pseudo-relevance feedback method to obtain scores for sentence and query semantic similarity is proposed [29]. In another study [27], a trained BERT model is utilized to generate character vectors that dynamically represent semantics based on character context. Consequently, it becomes evident that BERT-based models yield more accurate vector representations by considering both global and local word contexts [17].

To illustrate, one could consider an example using BERT embedding. For instance, given the sentence: "Saya suka makan nasi goreng (I like to eat fried rice)", contextual vector representations for each word in the sentence can be obtained. These vector representations reflect the meanings of the words based on the overall context of the sentence. For example, the vector representation for the word "Saya" (I) could be [0.12, 0.67, ..., 0.34], "suka" (like) could be [-0.29, 0.56, ..., 0.91], "makan" (eat) could be [0.76, 0.21, ..., 0.12], "nasi" (rice) could be [0.18, 0.62, ..., -0.06], and "goreng" (fried) could be [-0.42, 0.72, ..., 0.53]. These vector representations are obtained using a pre-trained BERT model with large-scale data. With these vector representations, various analyses and semantic operations can be performed. For instance, the cosine similarity between word vectors can be calculated, or the meanings of words in different sentences can be compared. For example, the vector representation of the word "suka" in the sentence "Saya suka makan nasi goreng" could be compared with another sentence like "Dia suka makan sushi (He likes to eat sushi)”. Through this comparison, differences and similarities in the meaning of the word "suka" in different contexts can be identified.

2.3 Hybrid-based approach

A hybrid-based or mixed-based approach to semantic expansion integrates multiple Query Expansion (QE) techniques. As demonstrated in previous studies, the amalgamation of both linguistic (lexicon-based) and semantic (ontology-based) strategies can enhance QE mechanisms, as corroborated by works [18, 19, 43]. Furthermore, the study [24] underscores the advantages of dovetailing lexicon and ontology techniques during the term expansion process. Lexicon techniques offer insights into correlated terms, while ontology techniques afford contextual understanding for each term.

A knowledge-based approach, such as utilizing a resource like WordNet, has been shown to assist in elucidating ambiguous terms in short texts [26]. By referring to the hierarchy of concepts and word relationships within WordNet, synonyms, antonyms, or other semantic associations for ambiguous terms can be identified. For example, if a word possesses multiple meanings, the knowledge base can facilitate the determination of the most appropriate context by considering associated synonyms.

Furthermore, approaches predicated on word embeddings, such as the employment of models like Word2Vec or BERT, have the capacity to capture the semantic connections between words in short texts. These models learn representations of words based on their distribution in extensive corpora. With these rich word representations, the textual context can be comprehended and semantic relationships between words discerned. For instance, if certain words frequently co-occur in texts containing hate speech, the word embedding model can detect this pattern and associate these words with the context of hate speech.

The relevance of these two techniques is magnified when applied to the classification of short texts and the detection of hate speech in the Indonesian language, a language abundant with ambiguous terms. By leveraging knowledge bases and word embedding models enriched with Indonesian data, the ability to classify short texts more accurately and identify hate speech more effectively can be augmented. This methodology aids in understanding the context and deeper meanings within the text, thereby making a significant contribution to the improvement of hate speech classification and detection in Indonesia.

3. SEMANTIC APPROACH TO EXPANSION OF SHORT TEXTS OF HATE SPEECH IN INDONESIAN

3.1 The challenge of detecting hate speech in Indonesian

The task of identifying hate speech within Indonesian social media discourse manifests as a considerable challenge, given the rampant multilingual communication. In the societal context of Indonesia, hate speech and harsh language, though understood in similar veins, exhibit distinct utilizations. Hate speech encompasses sentences, utterances, or writings laden with negative expressions, insults, or provocations aimed at individuals or groups, discriminating based on attributes such as race, religion, ethnicity, gender, or sexual orientation.

The primary intent of such hate speech is the propagation of hatred, incitement of hostility, or defamation of individuals or groups, often contravening legal frameworks and yielding deleterious societal impacts [5]. Conversely, the term 'rude words' delineates language characterized by impoliteness, vulgarity, or offensiveness in everyday parlance [44]. Profanity, typically, does not target specific groups based on protected attributes but refers to the indiscriminate use of impolite or abusive language.

Although the employment of profanity may not consistently intersect with illegality, it is ubiquitously regarded as ethically questionable or inappropriate within societal norms. Pertaining to the research under consideration, the constructed dataset incorporates both swear words and instances of hate speech, thus encompassing general profanity and specific instances of hate speech. These instances are defined within the confines of the legislative framework as outlined in the guidelines established by Komnas HAM [5].

3.2 Ambiguity issues and the need for disambiguation

Within the confines of the Indonesian language, hate speech is typically characterized as actions that insult specific individuals or groups, discriminating based on ethnicity, religion, race, or societal differences. Such forms of speech commonly incorporate offensive language capable of sparking negative emotional responses, with the potential to escalate societal conflicts to the point of genocide [4, 5]. However, it must be noted that offensive language does not consistently signify hate speech; it may occasionally be deployed humorously, albeit with the propensity to engender misunderstandings. Consequently, a detection system that accurately discriminates between hate speech and offensive language in Indonesian-language tweets is deemed necessary [45]. To enable precise identification of each word within short texts and to mitigate ambiguity in meaning, the implementation of semantic expansion emerges as a potent solution meriting additional exploration.

Consider the following illustration of hate speech: "Kamu adalah sampah (You are trash)". In this instance, the WordNet knowledge base might be utilized in the following ways:

Identification of synonyms: Possible synonyms for "trash" might include "waste" or "polluter".

Determination of the most suitable meaning: WordNet offers a comprehensive definition for "trash", enabling users to comprehend the word's context within hate speech.

Examination of word relations: WordNet elucidates the relationships between words, such as hypernyms (superordinate terms) and hyponyms (subordinate terms). Observing these relationships can aid users in understanding the nuanced meanings and broader context within the sentence.

3.3 Semantic expansion as a solution

The expansion of short texts for hate speech detection has emerged as a significant research interest, aiming to bolster the extraction of relevant information and facilitate a more comprehensive understanding of sentence context. Short texts often fail to provide users with expected information [3], leading to issues such as weak word correlation, skewed classification, suboptimal performance, and challenges to text classification algorithms. Therefore, the integration of semantic expansion exerts a significant influence on classification results [46]. Text extensions play an integral role in enhancing semantic contextual understanding through QE-based approaches.

For example, a hate speech detection system might be employed to analyze social media comments. A key challenge lies in identifying sentences containing hate speech while using synonyms or phrases not directly detectable. A comment on Twitter stating "Kamu memalukan (You are embarrassing)" exemplifies this challenge. The detection system may have been trained to recognize words like "shame" directly. However, hate speech often uses synonyms to dodge detection. Semantic expansion techniques can assist by broadening the detection scope through related synonyms, considering words such as "shame", "defamatory", or "damaging to reputation" as semantic extensions of "embarrassing". However, due to the presence of noise in captured documents and the heuristic-based nature of interpolation weights, errors may accumulate and impact final performance [47].

By employing semantic expansion techniques, such as incorporating synonyms or utilizing phrase construction, the hate speech detection system can bolster its ability to identify sentences containing hate speech that use synonyms or phrase constructions not directly detected. As a result, the system becomes more proficient at accurately detecting instances of hate speech and adeptly adapting to variations in language usage across distinct contexts.

3.4 Prior work on semantic expansion for Indonesian short texts

In addressing the prevalent challenges of data sparseness, noise, and inadequate information in short texts, a study [48] pioneered the development of an automatic sentiment analysis system for short informal Indonesian texts. This system leverages Naïve Bayes and Synonym-Based Feature Expansion, expanding words by utilizing the Kateglo thesaurus to ascertain multiple synonyms for each word in the original text, subsequently integrating them. Another research effort [47] applies the BM25 algorithm coupled with Word2vec as a Semantic Query Expansion method aimed at enhancing the search system model. However, this study reveals that the BM25 algorithm fails to locate relevant documents for terms bearing a semantic relationship with the initial query, resulting in the potential omission of pertinent documents from the user's search results.

Recognizing semantic relationships within multilabel data presents a formidable challenge. To address this, a semantic feature strategy based on Word2Vec [49]. This strategy involves transforming words into vector representations, facilitating the identification of semantic relationships by computing their distances using cosine similarity. Consider the sentence "a rock floats in a river". Post stemming, the sentence alters to "pumice in a river", revealing a clear difference in meaning between the two sentences.

To combat the issue of polysemy, the semantic expansion strategy in this context hinges on the Part-of-Speech (POS) approach. An additional research undertaking [50] directs its focus towards data expansion through a data augmentation-based approach to improve the detection of cyberbullying on social media. For the generation of coherent words from social media data, disambiguation of word meanings and the establishment of synonymous relations are performed using the WordNet lexical database. Data augmentation techniques can also be applied in alternative ways, as demonstrated by researchers [51, 52], who employ the back-translation method for augmentation. This technique generates new data mirroring the meaning of the original data, without increasing the dataset's size, but with variations from the original data. Based on these studies, it is apparent that text expansion grounded in the QE approach enhances recall and precision, particularly in addressing ambiguity issues in short hate speech texts within the Indonesian context.

In response to this challenge, we propose a hybrid semantic expansion framework for the detection of hate speech within Indonesian-language Twitter data. Initially, our framework capitalizes on the WordNet and Kateglo knowledge bases for query expansion. It further incorporates the Word2Vec and IndoBERT word insertion models, trained on an Indonesian language corpus, to discern semantic similarities. Lastly, deep learning models, specifically Convolutional Neural Networks (CNN), are deployed for the classification and detection processes. Through this multifaceted approach, our framework amplifies semantic understanding, captures contextual meaning, and ultimately determines whether a given short text constitutes hate speech.

4. A PROPOSED FRAMEWORK: SEMANTIC APPROACH BASED ON WORD EMBEDDING

The aim of our research is to address preceding challenges, including ambiguity, data dispersion, and polysemy, which impede hate speech detection. In this section, a hybrid approach, which amalgamates knowledge-based and embedding techniques for semantic expansion, is proposed. Each method brings unique advantages. The utilization of WordNet and Kateglo as knowledge bases aids in identifying lexical semantic relationships, thereby facilitating disambiguation [53]. Conversely, the integration of BERT embedding enables the capture of contextual semantics by generating meaning representations anchored on sentence structure [54]. To achieve comprehensive expansion and bolster semantic understanding, three BERT embedding models trained on the Indonesian language corpus are employed: IndoBERT, IndoBERT-Tweet, and BERT-multilingual. The Word2vec model, also trained on the Indonesian language, is included for experimental comparison. These techniques' integration allows for expansive word expansion, enhancing semantic understanding and aiding in the selection of the most fitting model for semantic broadening and hate speech detection.

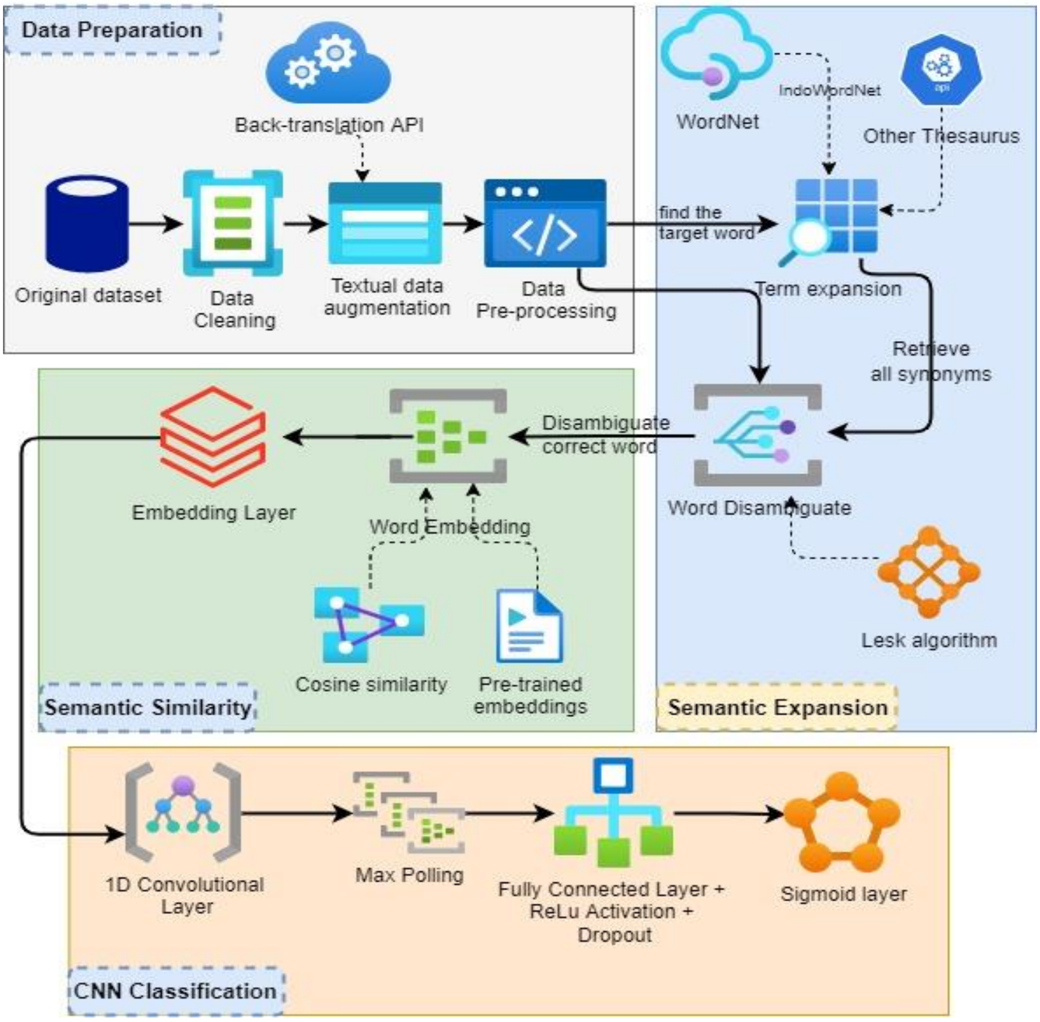
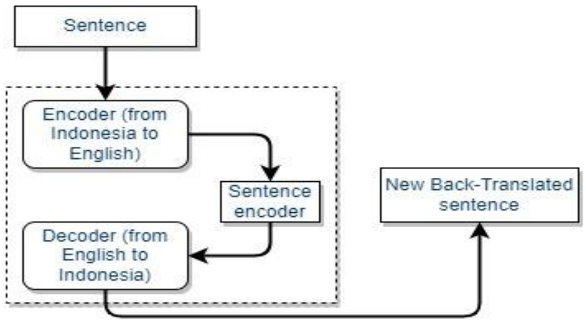


Figure 3. General framework for the process of text expansion based on semantics

In this section, a framework for semantic expansion and hate speech classification is presented, comprising four primary components: (1) data preparation, (2) semantic expansion, (3) semantic similarity, and (4) classification (refer to Figure 3 above). The subsequent discussion provides a detailed overview of these steps

4.1 Data preparation


Sentence
Encoder (from Indonesia to English)
Sentence encoder
Decoder (from English to Indonesia)
New Back-Translated sentence

Note: these cases can be extended to other languages until a suitable sentence structure is obtained

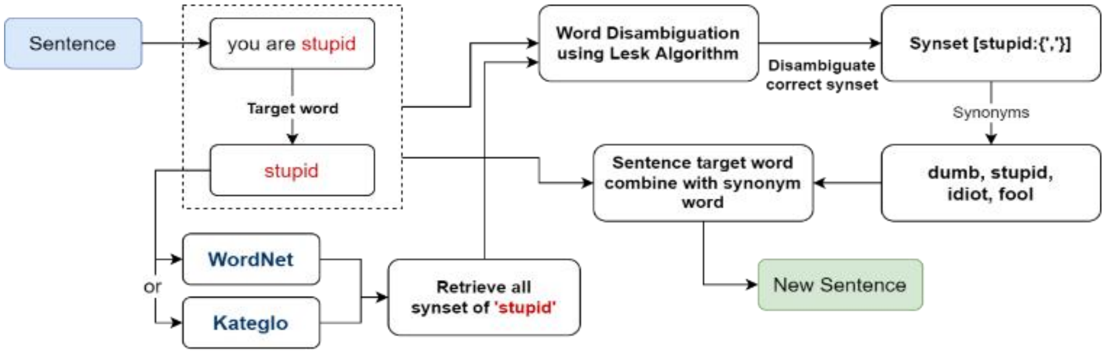
Figure 4. Data augmentation process using back-translation

Data was initially procured by scraping Twitter social media. To address the challenge posed by the unstructured grammar commonly found in Indonesian-language tweets, which often complicates comprehension, textual data augmentation techniques were employed. The objective was to modify the sentence structure to enhance clarity. For this purpose, back-translation was used for data augmentation and integrated into the dataset (refer to Figure 4). This approach, previously employed by researchers [51, 52], involves translating each sentence from Indonesian to English and back to Indonesian. For example, the sentence "kalian manusia bangsat tak butuh belas kasihan. Sialan (you bastards don't need pity. Damn it)" underwent the back-translation process, resulting in "Anda bajingan tidak perlu belas kasihan. Brengsek (You bastards don't need pity. Jerk)", which retains the original meaning. Once the sentence structure was enhanced, the data preprocessing stage was initiated.

4.2 Semantic expansion

This section elucidates the extraction and identification of word terms based on their utilization as verbs, adjectives, or nouns within sentences. The primary aim of this term expansion process is the generation of a comprehensive set of terms designed to significantly elevate the efficacy and precision of search results. In order to realize this, knowledge bases, such as WordNet and related thesauri, are harnessed, as they offer invaluable resources for user query expansion, thus enabling the capture of a wider range of terms. To address potential ambiguities in candidate word terms, the Lesk algorithm, a potent tool for disambiguation, is deployed. WordNet and Kateglo are instrumental in identifying potential terms for short word expansions, leveraging their vast word collections and complex semantic relationships. By subjecting each target word to verification via WordNet, its most suitable meaning within a specific context can be determined, guided by the Lesk algorithm.

To exemplify the expansion of short texts and the disambiguation process via the Lesk algorithm (refer to Figure 5), consider the sentence "Kamu sangat bodoh! (You are stupid!)". The word "bodoh" can have multiple meanings, each with its distinct semantic associations. Upon consulting WordNet, it is found that "bodoh" can imply a lack of intelligence or common sense, and it is associated with words such as "tolol," "idiot," and "bego". Subsequently, the Lesk algorithm is utilized to calculate the degree of overlap between the definitions of these words and the contextual words extracted from WordNet. A higher degree of overlap signifies a more accurate determination. For instance, in the context of the sentence, the meaning "not smart" is selected for "stupid", and the disambiguation results are integrated into the sentence as follows: "you are very stupid (not smart)".


Sentence
you are stupid
Target word
stupid
WordNet
or
Kateglo
Retrieve all synset of 'stupid'
Word Disambiguation using Lesk Algorithm
Disambiguate correct synset
Synset [stupid:{',')]
Synonyms
dumb, stupid, idiot, fool
Sentence target word combine with synonym word
New Sentence

*Note: **Kateglo** is an Indonesian dictionary, thesaurus and glossary*

Figure 5. Illustration of expanding short texts and word disambiguation using WordNet and the Lesk algorithm

4.3 Semantic similarity

In the subsequent phase, each word in the sentence undergoes a transformation into a multidimensional vector. Words with similar meanings or contexts are positioned in close geometric proximity within the embedding space. To gauge semantic similarity, a pretrained BERT model was utilized, and similarity scores were evaluated using cosine similarity, thereby identifying the most appropriate candidates for expansion terms [55]. Calculating the similarity between the original query or text and the candidate terms involved the use of the BERT model to secure vector representations for both text forms. Following this, the cosine similarity algorithm was applied to measure the cosine of the angle between the two vectors [56]. By determining the alignment of the vectors in the same direction, the cosine similarity algorithm gauges the degree of semantic similarity. A cosine similarity value nearing 1 indicates a higher degree of semantic similarity between the vectors.

For example, consider the query: "Kamu sangat bodoh! (You are stupid!)". The candidate terms generated are "tolol", "idiot", "bego". This process can be bifurcated into two stages: (1) the BERT model is used to generate embedding vectors for the query and the candidate terms, rendering the following results: (a) embedding vector for the query: [0.1, -0.6, 0.2], (b) embedding vector for "tolol": [0.2, -0.7, 0.1], (c) embedding vector for "idiot": [0.3, -0.4, 0.6], (d) embedding vector for " bego ": [-0.1, -0.8, 0.3]. (2) The cosine similarity between the embedding vector for the query and the vectors for the candidate terms is calculated using the cosine similarity algorithm, rendering the following results: (a) cosine similarity between the query vector and the term "tolol": 0.945, (b) cosine similarity between the query vector and the term " idiot": 0.857, (c) cosine similarity between the query vector and the term "bego": 0.762.

In this example, the candidate term "tolol" demonstrates the highest cosine similarity value, 0.945, indicating its close alignment with the hate speech expression "Kamu sangat bodoh!". This feature aids the hate speech detection system in identifying and expanding commonly used terms within the context of hate speech.

4.4 Classification

Upon completion of the preceding data processing, a binary classification of hate speech was performed. The original, pre-expanded dataset was randomly partitioned into three sections: 80% allocated for training, 10% for validation, and 10% for testing [45]. It is noteworthy that all experiments conducted within this study adhered to a similar test setup.

During the classification process, a Convolutional Neural Network (CNN) was utilized, with parameter settings adapted from the CNN architecture delineated in Kim's study [57]. The input layer represented a composite of words, harnessing the IndoBERT embedding with 300 embedding vectors for each word. A 1D convolution operation was employed with a kernel size of 3, followed by a maximum pooling operation on a feature map with a layer density of 50. To regulate the model, dropout was integrated into the penultimate layer. For a visual representation of the CNN architecture, refer to Figure 6.

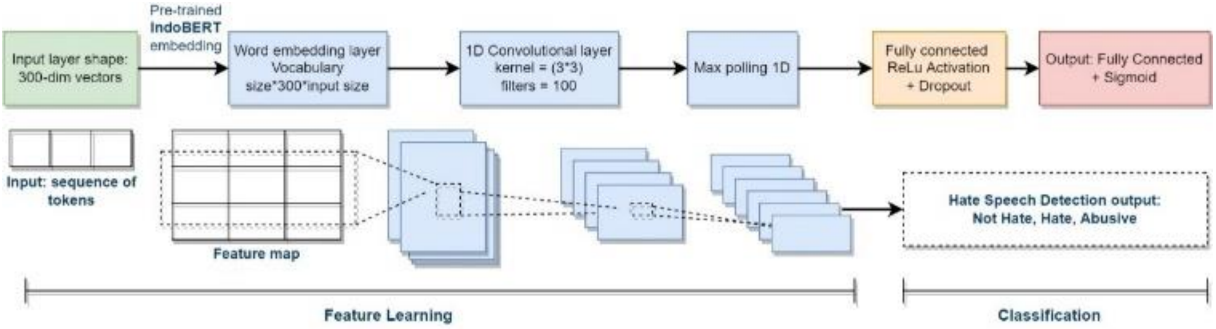

Pre-trained
IndoBERT
embedding
Input layer shape: 300-dim vectors
Word embedding layer Vocabulary size*300*input size
1D Convolutional layer kernel = (3*3) filters = 100
Max polling 1D
Fully connected ReLu Activation + Dropout
Output: Fully Connected + Sigmoid
Input: sequence of tokens
Feature map
Hate Speech Detection output: Not Hate, Hate, Abusive
Feature Learning
Classification

Figure 6. Proposed architecture for hate speech detection using CNN and IndoBERT

It is essential to acknowledge that while the proposed framework illustrates the potential of semantic expansion in understanding short texts, its limitations, which remain untested, must be recognized. These include: (1) reliance on limited resources dedicated to the Indonesian language for the knowledge base and embeddings used, (2) the resource-intensive nature and complexity of identifying optimal expansion candidates and generating extensible data, and (3) the potential for loss of meaning or context during the back-translation process between English and the original language, which could adversely impact performance.

Semantic broadening can substantially contribute to addressing challenges such as ambiguity, data dispersion, and polysemy, thereby enhancing the detection of hate speech. The selection of candidate terms during semantic expansion is reliant not only on their semantic similarity to the original query but also on their context and usage.

Consider the following illustration for insight into context and usage: Original query: "I hate him".

(1) Scenario without semantic expansion: If only the direct semantic equivalents of the original query are considered, candidate terms like "hate", "angry", or "resentful" might be considered. However, this approach fails to offer a precise understanding of the context and intensity of the hate speech involved.

(2) Scenario with semantic expansion: Through semantic expansion, we can explore broader meanings and contexts of the words used. For instance, synonyms or related words that carry more specificity in the context of hate speech can be explored. In this case, the resulting candidate terms may include "hates vehemently", "resents", or "feels intense dislike". By considering the context and usage of words, candidate terms that are more relevant and accurately represent the intensity of the hate speech can be identified.

In the given example, semantic expansion aids in selecting candidate terms that better align with the intensity and context of the expressed hate speech. As a result, this approach enhances hate speech detection by considering not only semantic similarities but also the contextual aspects and usage of terms within sentences.

5. CONCLUSIONS, CURRENT CHALLENGES, AND FUTURE DIRECTIONS

The research herein aimed to explore the application and efficacy of semantic expansion techniques in the classification of short Indonesian texts, specifically for hate speech detection. This investigation was guided by two key research questions: (1) How can semantic expansion techniques be applied to short text classification? and (2) What is the effectiveness of semantic expansion in mitigating term ambiguity in short Indonesian texts?

In addressing the first question, it was found that semantic expansion could be effectively applied to short text classification by enriching the semantic representation of words or phrases. This enhancement in semantic representation significantly improved the understanding of short text meanings, thus enabling classification algorithms to discern hate speech from non-offensive content more accurately. This approach was exemplified through a process of synonym replacement, whereby words or phrases were substituted with semantically comparable alternatives.

To answer the second research question, a hybrid approach was adopted, integrating knowledge-based techniques such as WordNet and BERT embeddings. These techniques were instrumental in identifying semantic relationships and capturing contextual nuances. Despite these advancements, it was acknowledged that semantic expansion alone cannot wholly eradicate ambiguity, particularly when contextual information is limited. Therefore, the incorporation of additional approaches, such as context analysis or advanced natural language processing techniques, was suggested to address term ambiguity more accurately.

The outcomes of the present study underscore the potential of semantic expansion methods in reducing term ambiguity within short Indonesian texts, thus enhancing the understanding of their intended meanings. However, despite the promising results, several limitations were identified that highlight future research opportunities in the application of semantic-based short text classification.

The first limitation pertains to data dissemination. The challenge of data dispersion in hate speech detection stems from the limited availability of relevant, well-labeled data sources, particularly for the Indonesian language with its varied dialects and writing styles. The dispersion of social media data across multiple platforms further complicates the process of data collection. Developing an effective semantic-based hate speech detection model requires a substantial and diverse dataset for training.

The second limitation concerns knowledge-based constraints. Detecting hate speech often necessitates an in-depth understanding of the cultural and social contexts in which the texts are situated. However, the available knowledge base may limit this understanding. For instance, in the case of Indonesian hate speech, specific terms or phrases may not be adequately documented in existing semantic knowledge bases such as WordNet, thereby restricting the accurate identification of relevant terms in the hate speech context.

The third limitation revolves around the issue of ambiguity in hate speech detection. Terms used in negative contexts can possess multiple meanings, making it challenging to determine their intended use. While an approach considering context and term usage could mitigate this challenge, the complexity of this task should not be underestimated.

The final limitation pertains to computational constraints. The detection of hate speech presents a complex, continually evolving challenge, with hate speech often conveyed indirectly through ambiguous sentence structures, parables, or figurative language. Corpus-based systems that employ computationally intensive calculations to extract latent semantics from training corpora further contribute to this complexity.

Despite these limitations, the proposed framework offers significant potential for advancing the semantic-based classification of short texts. Future research directions could include broadening data sources, improving knowledge bases, and developing methods to facilitate multilingual and cross-language expansion of short texts. These advancements would enhance the breadth of language coverage, deepen the understanding of cultural and social contexts in hate speech, and enable more effective detection and classification of hate speech across various languages.

By addressing these limitations, the proposed work can contribute to the advancement of semantic-based short text classification and improve the effectiveness of hate speech detection. This research will continue to inform strategies for combating hate speech, assisting content moderators and policy makers in identifying and mitigating potentially harmful online content.

REFERENCES

[1] Modha, S., Majumder, P., Mandl, T., Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber watchdog for surveillance. Expert Systems with Applications, 161: 113725. https://doi.org/10.1016/j.eswa.2020.113725

[2] Murugesan, S., Kaliyamurthie, K.P. (2023). A machine learning framework for automatic fake news detection in indian tamil news channels. Ingénierie des Systèmes d'Information, 28(1): 205-209. https://doi.org/10.18280/isi.280123

[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] 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. http://dx.doi.org/10.18653/v1/W19-3506

[5] Manusia, K.N.H.A. (2015). Buku Saku Penanganan Ujaran Kebencian (Hate Speech). Jakarta: Komnas HAM.

[6] 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, Springer London, pp. 61-69. https://doi.org/10.1007/978-1-4471-2099-5_7

[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] Pennington, J., Socher, R., Manning, C.D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. http://dx.doi.org/10.3115/v1/D14-1162

[9] Bengio, Y., Ducharme, R., Vincent, P. (2000). A neural probabilistic language model. In: Holmes, D.E., Jain, L.C. (eds) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol 194. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33486-6_6

[10] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12: 2493-2537.

[11] Kuzi, S., Shtok, A., Kurland, O. (2016). Query expansion

using word embeddings. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 1929-1932. https://doi.org/10.1145/2983323.2983876

[12] Diaz, F., Mitra, B., Craswell, N. (2016). Query expansion with locally-trained word embeddings. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 367-377. http://dx.doi.org/10.18653/v1/P16-1035

[13] Le, Q., Mikolov, T. (2014). Distributed representations of sentences and documents. In International Conference on Machine Learning, pp. 1188-1196.

[14] Liang, W., Feng, R., Liu, X., Li, Y., Zhang, X. (2018). GLTM: A global and local word embedding-based topic model for short texts. IEEE Access, 6: 43612-43621. https://doi.org/10.1109/ACCESS.2018.2863260

[15] 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

[16] Yeke, D. (2020). Improving document ranking with query expansion based on bert word embeddings. Master's thesis, Middle East Technical University.

[17] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 304: 114135. https://doi.org/10.1016/j.psychres.2021.114135

[18] Devi, M.U., Gandhi, G.M. (2015). Wordnet and ontology-based query expansion for semantic information retrieval in sports domain. Journal of Computer Science, 11(2): 361-371. http://dx.doi.org/10.3844/jcssp.2015.361.371

[19] 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

[20] Dahir, S., El Qadi, A. (2021). A query expansion method based on topic modeling and DBpedia features. International Journal of Information Management Data Insights, 1(2): 100043. https://doi.org/10.1016/j.jjimei.2021.100043

[21] Xu, J., Cai, Y., Wu, X., Lei, X., Huang, Q., Leung, H.F., Li, Q. (2020). Incorporating context-relevant concepts into convolutional neural networks for short text classification. Neurocomputing, 386: 42-53. https://doi.org/10.1016/j.neucom.2019.08.080

[22] Bhogal, J., MacFarlane, A., Smith, P. (2007). A review of ontology-based query expansion. Information Processing & Management, 43(4): 866-886. https://doi.org/10.1016/j.ipm.2006.09.003

[23] Li, Y., Luk, W.P.R., Ho, K.S.E., Chung, F.L.K. (2007). Improving weak ad-hoc queries using wikipedia asexternal corpus. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 797-798. https://doi.org/10.1145/1277741.1277914

[24] Storey, V.C., Burton-Jones, A., Sugumaran, V., Purao, S. (2008). CONQUER: A methodology for context-aware query processing on the World Wide Web. Information Systems Research, 19(1): 3-25. https://doi.org/10.1287/isre.1070.0140

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

[26] Pinto, F.J., Martinez, A.F., Perez-Sanjulian, C.F. (2008). Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet. International Journal of Computer Applications in Technology, 33(4): 271-279. https://doi.org/10.1504/IJCAT.2008.022422

[27] Wang, L., Niu, J., Yu, S. (2019). SentiDiff: Combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 32(10): 2026-2039. https://doi.org/10.1109/TKDE.2019.2913641

[28] Pan, M., Wang, J., Huang, J.X., Huang, A.J., Chen, Q., Chen, J. (2022). A probabilistic framework for integrating sentence-level semantics via BERT into pseudo-relevance feedback. Information Processing & Management, 59(1): 102734. https://doi.org/10.1016/j.ipm.2021.102734

[29] Hsu, M.H., Tsai, M.F., Chen, H.H. (2006). Query expansion with conceptnet and wordnet: An intrinsic comparison. In: Ng, H.T., Leong, MK., Kan, MY., Ji, D. (eds) Information Retrieval Technology. AIRS 2006. Lecture Notes in Computer Science, vol 4182. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11880592_1

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

[31] Altınel, B., Ganiz, M.C. (2018). Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6): 1129-1153. https://doi.org/10.1016/j.ipm.2018.08.001

[32] Harman, D. (1992). Relevance feedback revisited. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1-10. https://doi.org/10.1145/133160.133167

[33] Salton, G., Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4): 288-297. https://doi.org/10.1002/(SICI)1097-4571(199006)41:4%3C288::AID-ASI8%3E3.0.CO;2-H

[34] Singh, J., Sharan, A. (2017). A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Computing and Applications, 28: 2557-2580. https://doi.org/10.1007/s00521-016-2207-x

[35] Raza, M.A., Mokhtar, R., Ahmad, N., Pasha, M., Pasha, U. (2019). A taxonomy and survey of semantic approaches for query expansion. IEEE Access, 7: 17823-17833. https://doi.org/10.1109/ACCESS.2019.2894679

[36] AlMasri, M., Berrut, C., Chevallet, J.P. (2016). A comparison of deep learning-based query expansion with pseudo-relevance feedback and mutual information. In Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38, Springer

International Publishing, pp. 709-715. https://doi.org/10.1007/978-3-319-30671-1_57

[37] Xu, Y., Jones, G.J., Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59-66. https://doi.org/10.1145/1571941.1571954

[38] Noor, N.H.B.M., Sapuan, S., Bond, F. (2011). Creating the open wordnet bahasa. In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, pp. 255-264.

[39] ALMasri, M., Berrut, C., Chevallet, J.P. (2013). Wikipedia-based semantic query enrichment. In Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, pp. 5-8. https://doi.org/10.1145/2513204.2513209

[40] Roy, D., Paul, D., Mitra, M., Garain, U. (2016). Using word embeddings for automatic query expansion. arXiv preprint arXiv: 1606.07608. https://doi.org/10.48550/arXiv.1606.07608

[41] Baly, R., Hajj, H., Habash, N., Shaban, K.B., El-Hajj, W. (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16(4): 1-21. https://doi.org/10.1145/3086576

[42] Wang, Z., Huang, Z., Gao, J. (2020). Chinese text classification method based on BERT word embedding. In Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, pp. 66-71. https://doi.org/10.1145/3395260.3395273

[43] Jain, S., Seeja, K.R., Jindal, R. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. International Journal of Information Management Data Insights, 1(1): 100009. https://doi.org/10.1016/j.jjimei.2021.100009

[44] Rasel, R.I., Sultana, N., Akhter, S., Meesad, P. (2018). Detection of cyber-aggressive comments on social media networks: A machine learning and text mining approach. In Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval, pp. 37-41. https://doi.org/10.1145/3278293.3278303

[45] 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

[46] 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

[47] Purnama, A.R.G., Yulita, I.N., Helen, A. (2021). Search system for translation of Al-Qur'an verses in Indonesian using BM25 and semantic query expansion. In 2021 International Conference on Artificial Intelligence and Big Data Analytics, Bandung, Indonesia, pp. 1-7. https://doi.org/10.1109/ICAIBDA53487.2021.9689757

[48] Fauzi, M.A., Afirianto, T. (2018). Improving sentiment analysis of short informal Indonesian product reviews using synonym-based feature expansion. Telkomnika (Telecommunication Computing Electronics and Control), 16(3): 1345-1350. http://doi.org/10.12928/telkomnika.v16i3.7751

[49] 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

[50] 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.

[51] 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

[52] Sennrich, R., Haddow, B., Birch, A. (2015). Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709. https://doi.org/10.48550/arXiv.1511.06709

[53] Greenberg, J. (2001). Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. Journal of the American Society for information science and Technology, 52(6): 487-498. https://doi.org/10.1002/asi.1093

[54] Gupta, S., Lakra, S., Kaur, M. (2020). Study on bert model for hate speech detection. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, pp. 1-8. https://doi.org/10.1109/ICECA49313.2020.9297560

[55] Chandrasekaran, D., Mago, V. (2021). Evolution of semantic similarity—a survey. ACM Computing Surveys (CSUR), 54(2): 1-37. https://doi.org/10.1145/3440755

[56] El Ghali, B., El Qadi, A., Ouadou, M., Aboutajdine, D. (2015). Context-based query expansion method for short queries using latent semantic analyses. In: Bouajjani, A., Fauconnier, H. (eds) Networked Systems. NETYS 2015. Lecture Notes in Computer Science, vol. 9466. Springer, Cham. https://doi.org/10.1007/978-3-319-26850-7_33

[57] Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751. https://doi.org/10.3115/v1/d14-1181