

Enhancing hate speech detection in Indonesian using abusive words lexicon

Endang Wahyu Pamungkas^{1,4}, Dian Purworini^{2,4}, Divi Galih Prasetyo Putri³, Sohail Akhtar⁵

¹Department of Informatics Engineering, Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

²Department of Communication Science, Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

³Department of Electrical Engineering and Informatics, Vocational College, Universitas Gadjah Mada, Yogyakarta, Indonesia

⁴Social Informatics Research Center, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

⁵Department of Computer Science, Faculty of Engineering Science, Bahria University, Islamabad, Pakistan

Article Info

Article history:

Received Oct 13, 2023

Revised Nov 11, 2023

Accepted Nov 15, 2023

Keywords:

Abusive language
Hate speech detection
Low-resource languages
Machine learning
Social media

ABSTRACT

Hate speech is a major challenge in Indonesia, a diverse country with multiple languages and a dynamic online landscape. This research explores the phenomenon of hate speech and its detection, particularly in language contexts with limited resources. We introduce a new abusive words lexicon, created by collecting words from various sources, adapted for Indonesian, Javanese and Sundanese. Our study investigates the practical implementation of this lexicon. We conducted extensive experiments using different datasets and machine learning models, aiming to improve hate speech detection. The results consistently show a positive impact of the lexicon, which significantly improves detection, especially in languages with fewer resources. But this research paves the way for further exploration. The lexicon can be expanded, broadening its scope. Additionally, we suggest investigating more sophisticated models, such as transformer-based models, to more effectively detect hate speech. In a world where hate speech is a growing problem, our research provides valuable insights and tools to combat it effectively in Indonesia and other countries.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Endang Wahyu Pamungkas

Department of Informatics Engineering, Universitas Muhammadiyah Surakarta

Surakarta, Central Java, Indonesia

Email: ewp123@ums.ac.id

1. INTRODUCTION

The rise of microblogging platforms has transformed how people communicate globally, providing a quick and far-reaching way for individuals to express their thoughts. Unfortunately, this digital landscape has also become a breeding ground for the rapid spread of harmful and aggressive content. This is made possible by the easy accessibility of these platforms, user anonymity, and people's tendency to assert their opinions vigorously. Hate speech, a subject of academic investigation, includes various forms of communication that express strong hostility or contempt towards individuals or groups based on characteristics like race, ethnicity, gender, sexual orientation, religion, or disability [1]. With the growing influence of social computing, online interactions have gained prominence, especially on social media and chat forums. While the legal definitions of hate speech may vary from one country to another, they all generally involve any communication that displays hatred or insults aimed at individuals or groups due to their specific attributes. Recognizing the harmful

impact of hate speech on online social media, numerous studies have been conducted to understand its negative effects. Hate speech detection plays a crucial role in this effort, utilizing advanced techniques in natural language processing and machine learning to automatically spot instances of hate speech. The ultimate goal is to identify and remove harmful content from online platforms and social media, creating a safer and more inclusive digital space for all users [2]. The issue of hate speech has become a significant concern in today's social media landscape. In some extreme instances, social media platforms can be exploited by hate groups to disseminate messages of animosity towards specific targets, potentially inciting dangerous criminal activities. Consequently, there is an urgent need to proactively curb and promptly address the proliferation of hate speech online, as unchecked, it may culminate in severe real-world crimes. The continually escalating user base on social media platforms has rendered manual content monitoring an arduous and non-scalable task. Consequently, various research endeavors have embarked on the development of automated hate speech detection models, employing diverse methodologies, ranging from traditional machine learning techniques [3], [4] to the utilization of deep learning algorithms [5], [6]. Nonetheless, the automated identification of hate speech in languages with limited linguistic resources presents a formidable challenge [7]-[9]. A case in point is the tragic escalation of violence against the Rohingya Muslim minority in Myanmar, which was exacerbated by the inability to effectively counteract the surge in hate speech on the social media platform Facebook, primarily due to the intricacies of automatically processing Burmese language texts. A similar challenge is evident in the case of Indonesian, where the availability of language resources remains severely constrained.

According to data sourced from the Criminal Investigation Agency of the Indonesian National Police, there were 143 documented cases of cybercrimes involving hate speech in the year 2015, which surged to 199 cases in the subsequent year, 2016. However, it is imperative to acknowledge that these statistics solely encompass instances of hate speech that were formally reported and classified as criminal offenses. Unquestionably, there exists a multitude of unreported cases of hate speech proliferating across various social media platforms. The challenge of hate speech in Indonesia is deeply rooted in its complex socio-cultural landscape. Indonesia is not only the world's most populous muslim-majority nation but also boasts remarkable diversity in terms of ethnicity, religion, and culture. This diversity, while a source of strength, can also be a breeding ground for intergroup tensions and hate speech. Various regions and communities in Indonesia have witnessed historical and socio-political conflicts that have left scars and, at times, perpetuated biases and prejudices. Furthermore, recent reports underscore the remarkable linguistic diversity within Indonesia, with 718 distinct local languages spoken by various regions and tribes. This rich linguistic tapestry presents a formidable challenge when processing Indonesian social media data, as interactions frequently involve a blend of local languages and the national language, Indonesian, both in everyday communication and online discourse [10]. The inherently informal nature of social media data further compounds the complexity of delineating the boundaries of hate speech within this context. Hence, the development of a model capable of effectively detecting hate speech within code-mixed data becomes of paramount importance. Recent studies have highlighted the formidable nature of this task, as identifying offensive language within datasets that incorporate multiple languages presents a daunting challenge [11], [12]. The use of offensive terms is heavily influenced by individual cultural nuances and exhibits significant variations from one language to another. Moreover, resource constraints, particularly concerning minority languages, add an additional layer of complexity to this endeavor [13].

According to statista, the number of Twitter users in Indonesia has soared to nearly 240 million, securing the country's fifth-place ranking globally in terms of Twitter users. This significant online presence highlights the potential for the proliferation of hate speech in the digital sphere. The Indonesian government has taken steps to regulate hate speech since 2008, as delineated in the Law of Information and Electronic Transaction (UU ITE). The *Kepolisian Republik Indonesia*, or the Indonesian Police Department, has further articulated regulations to combat hate speech, recognizing its capacity to inflict harm not only on individual victims but also on society as a whole. It is intriguing to note that the majority of hate speech incidents on Indonesian social media are triggered by political events, particularly during elections.

However, research into the detection of hate speech on Indonesian social media remains relatively nascent [14], [15], resulting in limited availability of linguistic resources and corpora. Most studies have concentrated on automating the detection of hate speech expressions from social media data. An early study by Alfina *et al.* [16] marked one of the pioneering efforts in hate speech detection within the Indonesian social media landscape, with a specific focus on the Twitter platform. This work introduced a unique dataset compiled from Twitter, meticulously annotated with two labels: hate speech and non-hate speech. Another significant contribution comes from Ibrohim and Budi [17], who developed a more comprehensive hate speech dataset.

This dataset not only encompasses binary classification (hate speech versus non-hate speech) but also provides annotations based on multiple categories, including the target of the hate speech, its category, and the degree of hatefulness. More recent endeavors in hate speech detection in Indonesia have gravitated towards the adoption of contemporary technologies, such as neural-based and transform-based models [18], [19]. These studies represent a concerted effort to enhance the precision and effectiveness of hate speech detection in the dynamic Indonesian social media landscape, aligning with the evolving landscape of online discourse.

In this study, our primary focus is on harnessing external knowledge derived from an abusive word lexicon to enhance the automated detection of hate speech, particularly within the context of low-resource Indonesian languages. By capitalizing on our understanding of offensive terminology in the Indonesian language, we anticipate a notable enhancement in the performance of hate speech detection algorithms. The aforementioned lexicon will be seamlessly integrated into a dedicated machine learning model designed specifically for identifying hate speech within low resources language settings. To realize this objective, we will explore the utility of several traditional machine learning models. In addition to these, we will also evaluate the effectiveness of two recurrent neural network (RNN)-based models: the gated recurrent unit (GRU) and the long short-term memory (LSTM). Our experimental investigation will encompass the utilization of four of the most widely recognized hate speech datasets in Bahasa Indonesia. This comprehensive approach aims to furnish a comprehensive understanding of the performance and adaptability of our proposed models in the real-world context of hate speech detection within Indonesian language resources. Specifically, we address the following research questions.

RQ1: What is an effective methodology for constructing an abusive word lexicon in Indonesian? To tackle this question, we propose the creation of a novel abusive lexicon through a comprehensive approach that involves sourcing words from various references, including translation, followed by a meticulous manual verification process.

RQ2: How can we effectively integrate external knowledge from a lexicon into machine learning models for the purpose of hate speech detection in Indonesian? Our investigation centers on assessing the influence of an abusive word lexicon on knowledge transfer within the context of low-resource Indonesian languages. We accomplish this by incorporating lexicon-derived information as a feature within the feature matrix representation.

Based on these research questions, we propose two pivotal contributions in this work which can be summarized as following: i). We present novel abusive word lexicons in three distinct languages, namely Indonesian, Javanese, and Sundanese. These lexicons are meticulously constructed based on available resources, supplemented by human manual verification. ii). We advocate for the integration of additional features from the abusive word lexicon into several machine learning models for enhanced hate speech detection. Our proposed approach is evaluated using various benchmark datasets.

This article comprises various sections, commencing with an introduction section 1. Followed by a review of relevant prior research section 2. Section 3 describes the adopted method in this study, which divided into two subsections. The approach for constructing the lexicon is outlined in subsection 3.1. The experimental procedures will be presented in subsection 3.2. The analysis and discussion of the Results will be presented in the section 4. Finally, section 5 will provide a summary of the paper.

2. RELATED WORKS

Numerous research effort have delved to deal with hate speech in Bahasa Indonesia. Predominantly, the available datasets for identifying hate speech in Indonesia are obtained from social media platforms like Twitter [16], [17], Facebook [20], and Instagram [21], with Twitter emerging as the prevailing data source. This prevalence could be attributed to the relative ease of sample acquisition through Twitter's accessible public API and the platform's comparatively lenient data-sharing policies. Several models has been applied to address the challenge of detecting hate speech in the Indonesian setting. Nevertheless, many studies have leaned towards conventional models [16], [17], including logistic regression, support vector machines (SVM), naive bayes (NB), and random forest (RF), to address this task. These models have been trained using several feature representations, encompassing term frequency-inverse document frequency (TF-IDF), bag of words, and word vectors derived from pre-existing language representations. While these traditional models have shown utility, recent strides in deep learning have catapulted them to the forefront, with current deep learning architectures consistently yielding more competitive results than their counterparts [22]. Notably, there has

been a surge of interest in employing transformer-based models to autonomously identify hate speech within Indonesian datasets, reflecting a shift towards harnessing advanced neural network architectures for this purpose [5], [19], [23], [24].

Utilizing lexicons as supplementary resources is not a novel concept in the realm of hate speech detection. The underlying idea is that lexicons can provide valuable information to machine learning models, augmenting their ability to classify data as hate speech or not, particularly in scenarios involving languages with limited resources. Chiril *et al.* [25] introduced an approach that incorporates various affective lexicons, including SenticNet [26], EmoSenticNet [27], and HurtLex [28], as emotional features to enhance hate speech detection across multiple datasets. This work posits that the emotional state of the author or speaker holds significance in disambiguating the context of a given utterance. The study demonstrated that the inclusion of emotional features indeed improved the model's performance in detecting hate speech. Furthermore, Pamungkas and Patti [29] proposed the utilization of the multilingual hate lexicon, HurtLex, to address domain-shift and language-shift challenges within the context of abusive language detection across various domains and languages. This research revealed that the additional features derived from HurtLex served as a bridging mechanism for transferring knowledge in cross-domain and cross-lingual hate speech detection tasks. Lastly, Koufakou *et al.* [30] presented a study that integrated state-of-the-art BERT models with supplementary knowledge sourced from HurtLex. The outcomes of this investigation demonstrated the efficacy of HurtLex features in enhancing model performance. Collectively, these studies underscore the potency of leveraging lexicons as complementary resources to bolster hate speech detection capabilities, showcasing the potential for improvement in various aspects of the task.

3. METHOD

To fulfill the aims of this study, we put forward a series of methodological processes as depicted in Figure 1. This research can be compartmentalized into two fundamental components. The initial phase centers on the development of a novel lexicon of offensive language, which draws upon existing linguistic resources. Subsequently, in the second phase, we leverage this newly constructed lexicon to augment the feature set used for enhancing the automated identification of hate speech across a range of benchmark datasets. The following sections provide a comprehensive exploration of these two pivotal phases.

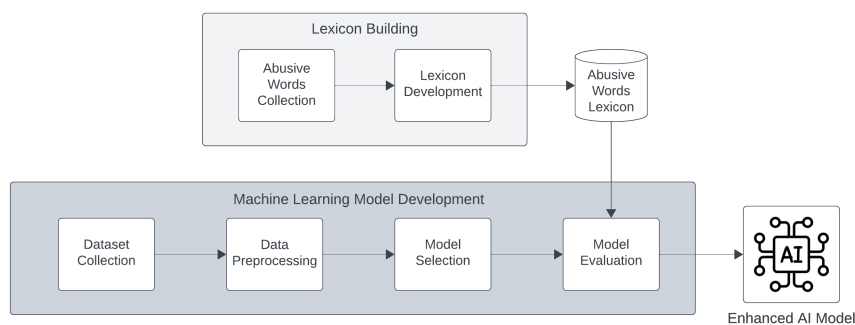


Figure 1. Research methodology

3.1. Development of abusive words lexicon

Abusive word lexicons are invaluable resources, consisting of collections of words and phrases employed to spot offensive, abusive, or potentially hateful content. These lexicons play a pivotal role in the realm of hate speech detection and related research, serving as reference guides to identify and filter out objectionable materials. One prominent example is hatebase an extensive compilation of offensive words in various languages, complete with contextual categorizations. Another noteworthy contribution comes from [31], who introduced a lexicon encompassing a roster of abusive terms. Empirical evidence substantiates the beneficial impact of this lexicon in effectively identifying abusive language across diverse domains. In addition to the previously mentioned lexicons, there exists HurtLex [28], a multilingual hate lexicon painstakingly translated from Italian into 53 different languages. HurtLex has showcased its effectiveness in bolstering hate speech detection, particularly within languages possessing limited linguistic resources [5], [29]. These lexicons stand as

essential tools in the pursuit of enhancing the accuracy and efficacy of hate speech detection methods across varying linguistic landscapes. To the best of my knowledge, there is limited existing research concerning the creation of lexicons specifically tailored for abusive language in the Indonesian context. Nevertheless, there is an iteration of HurtLex that has been semi-automatically translated into several languages, including Indonesian, using the BabelNet resource [32]. The efficacy of this Indonesian version of HurtLex has been demonstrated in mitigating the language shift challenge inherent in hate speech detection, particularly in the context of zero-shot learning [5]. Additionally, we have encountered a study that centers on the development of an abusive language lexicon encompassing the four most widely spoken local languages in Indonesia: Betawi, Madurese, Sundanese, and Javanese [33]. Upon examination, it becomes apparent that the lexicon generated in this study predominantly consists of explicit expressions of anger. In contrast, HurtLex encompasses a more extensive spectrum of implicit abusive terms. These two lexicons will serve as the foundational resources underpinning the current research. To answer our first research question, we propose to build a novel abusive words lexicon in which cover three different most used languages in Indonesian including Indonesian, Javanese, and Sundanese. In Figure 2, we present an overview of the planned lexicon development process for our experiment. As previously mentioned, the construction of this lexicon will draw upon three primary sources: HurtLex [28], the abusive word lexicon derived from the research conducted by [33], and the lexicon from [17]. To commence this procedure, we will translate HurtLex from Indonesian to Javanese and Sundanese languages using the Google Translate tool. This translation step will result in two distinct versions of HurtLex in both Javanese and Sundanese languages. Subsequently, these translated lexicons will be integrated with the lexicon instances obtained from [33] for each respective language. Simultaneously, the original Indonesian collection will be directly merged with the lexicon from [17]). Following the merging phase, we will execute a deduplication process to eliminate any redundant words. The resulting expanded lexicons will comprise 575 words for the Javanese language, 526 words for the Sundanese language, and 577 words for the Indonesian language. A comprehensive breakdown of the instance count at each stage of the lexicon development process, from its inception to the final compilation, is provided in Figure 2. Upon our initial examination, we noted that the lexicons from [33], [17] predominantly feature explicit expletive words, indicating a high degree of abusiveness. In contrast, the original HurtLex lexicon consists of implicit abusive terms. Consequently, the resulting lexicon from this study will encompass both implicit and explicit forms of hate speech, thereby enhancing its effectiveness for detecting hate speech within a broader contextual scope. Figure 3 depicts example of abusive words in Indonesian which also content of the lexicon.

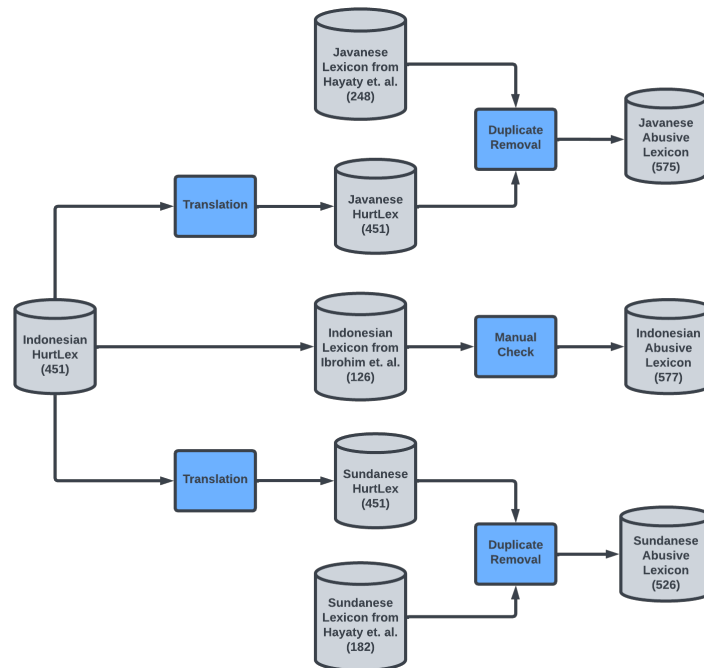


Figure 2. Lexicon building process

sites included detik, kaskus, and kompas. The annotation process engaged ten distinct annotators, with each annotator assigning one of three labels to each comment: “not abusive” (1), “abusive but not offensive” (2), and “abusive and offensive” (3). For the purpose of our research, we will consolidate the first and second classes into the category of “not hate speech,” while the third class will be categorized as “hate speech.”

3.2.2. Experimental settings

In preparation for our experimentation, we commence by partitioning the dataset into two segments: the training set and the test set. For all dataset collections, we allocate 70% of the data for training and reserve 30% for testing purposes. As part of our preliminary data processing procedures, we normalize the text by converting it to lowercase, and we standardize mentions (@) by replacing them with the term “USER”. Given the central focus of our research, which revolves around leveraging the lexicon to facilitate knowledge transfer and overcome the low resource language data, our methodology involves the integration of the lexicon as a feature within our machine learning model. This approach encompasses the utilization of a combination of traditional machine learning classifiers alongside deep learning models, specifically the LSTM and GRU. Subsequent sections of this work will provide an in-depth elucidation of the implemented models, as well as a detailed discussion of how we intend to incorporate the information derived from the lexicon.

Deep learning models: In the realm of deep learning methodologies, we incorporate two variations of RNN, specifically, the LSTM model [37] and the GRU model [38]. The application of this model in Bahasa Indonesia has been recognized beneficial [39], [40]. The architectural blueprint consists of multiple tiers, initiating with a 300-dimensional embedding layer. The embedding layer serves as the input for the ensuing LSTM or GRU network, which is composed of 64 units. This is succeeded by a dense layer featuring 16 units, employing the rectified linear unit (ReLU) activation function. The final prediction layer incorporates a dense layer with a sigmoid activation function. To optimize performance, we adapt the architecture to accommodate varying batch sizes (16, 32, 64) and a range of epoch values (1-5). Furthermore, to explore the implications of incorporating the lexicon, we integrate the same features employed in the conventional model approach. These supplementary features are concatenated within a dense layer after the LSTM or GRU network, allowing the model to harness the combined information derived from both the language-specific data and the lexicon-based features. The illustration of the proposed deep learning architecture can be seen in Figure 4.

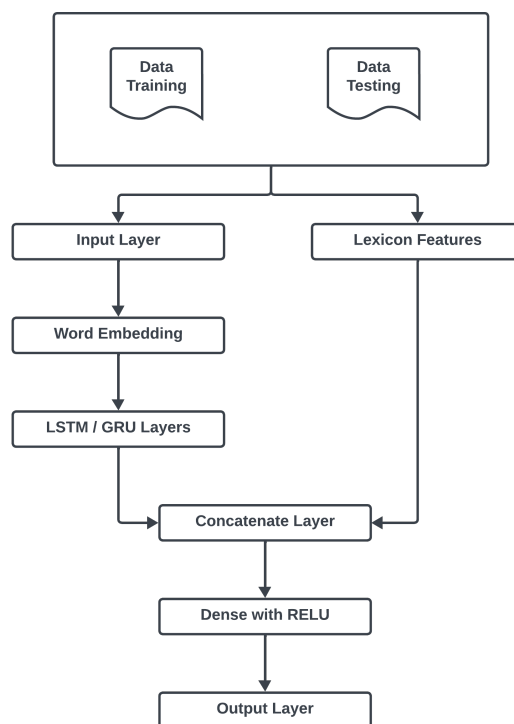


Figure 4. Deep learning architecture for lexicon feature injection

Traditional models: In this research endeavor, we will employ four distinctive classical machine learning classifier models: the linear support vector classifier (LSVC), decision tree (DT), logistic regression (LR), and RF classifier. To implement these models, we will harness the Scikit-learn library, utilizing the default parameter configurations provided by Scikit-learn, without engaging in hyperparameter optimization. As a fundamental benchmark system, our feature representation will be exclusively composed of bag-of-words. This representation will encompass unigram, bigram, and trigram representations to train our models. To assess the influence of lexicon integration, we will introduce it as an additional feature. This lexicon-derived feature will quantitatively reflect the extent of word matches between the tweet text and the lexicon entries.

4. RESULTS AND ANALYSIS

The outcomes of our experiments are presented in Tables 1-5, delineating a comprehensive comparison among various experimental scenarios. In the initial scenario (basic models), we implement a fundamental machine learning model devoid of supplementary features from the lexicon. In the subsequent scenario, we augment the basic model with supplementary features sourced from the lexicon, encompassing Indonesian, Javanese, or Sundanese lexicons (please refer to the lexicon development in section 3.1.). The primary objective underlying these two scenarios is to address our research inquiry regarding whether the added features from the abusive word lexicon can enhance the performance of machine learning models in hate speech detection within the Indonesian context. Through a meticulous analysis of these facets, we aim to acquire a profound comprehension of the efficacy of lexicon-based features and the distinctive characteristics of hate speech prevalent in the Indonesian social media landscape. We evaluate the performance of the machine learning models based on three different evaluation metrics including macro- F score (macro- F), accuracy (Acc), and area under the curve (AUC) score.

Our study uncovers a pivotal finding that highlights the consistent enhancement of machine learning model performance across diverse datasets. The incorporation of additional features from the proposed abusive word lexicon demonstrates remarkable improvements, particularly noteworthy in languages marked by limited linguistic resources. The Asti *et al.* [34] dataset exemplifies this scenario, especially in the case of Javanese and Sundanese languages (see Table 3 and Table 4). Across various machine learning models, including both traditional and deep learning architectures, the observed improvements are consistent. This underscores the crucial role played by external knowledge provided by the lexicon in effectively mitigating challenges associated with low-resource languages. The robust performance improvement in these low-resource languages suggests that the lexicon contributes valuable insights into the nuanced nature of hate speech within such linguistic contexts. The lexicon's ability to capture and represent abusive language specific to Indonesian, Javanese, and Sundanese contributes significantly to the models' heightened discriminatory power. The lexicon serves as a bridge, offering contextual understanding and aiding the models in effectively discerning hate speech instances in scenarios where linguistic resources are scarce.

One noteworthy observation is that traditional models outperformed their deep learning counterparts in datasets with a restricted number of instances. This outcome highlights the advantage of traditional machine learning approaches when data scarcity prevails. In such cases, deep learning models may struggle to generalize effectively due to limited data for parameter estimation. In contrast, traditional models can provide robust results with smaller datasets. However, the deep learning models displayed their strengths when confronted with more substantial datasets, as demonstrated by the datasets provided by Ibrohim and Budi [17] and Asti *et al.* [34]. In these scenarios, deep learning's capacity to learn intricate patterns from extensive data sources became evident, resulting in highly competitive results. This reinforces the idea that the choice between traditional and deep learning models should be influenced by the dataset's size and complexity. One intriguing aspect of our study is the underperformance of our implemented models in the Desrul and Romadhony [36] dataset. We attribute this outcome to the dataset's distinct annotation scheme and the associated challenges it introduces. The unique nature of the Desrul and Romadhony [36] dataset presents a compelling avenue for future research, highlighting the need for models that can adapt to diverse annotation styles and labeling intricacies.

For further justification and in-depth analysis, we have incorporated the best-performing models from each corresponding dataset's original studies. However, we did not include a comparison with the Desrul and Romadhony [36] dataset due to its distinct annotation scheme, rendering a meaningful comparison unfeasible. Our comparative analysis indicates that all of our models, including the baseline models, consistently outperform the models proposed in the original work of each respective dataset. Notably, in the

Ibrohim and Budi [17] dataset, our best-performing model achieved an accuracy of 0.849, while the original work achieved an accuracy of 0.735. In the case of the Pratiwi *et al.* [35] dataset, our top model achieved an *F*-score of 0.750, in contrast to the original models which scored 0.657. In both the Asti *et al.* [34] Javanese and Sundanese datasets, our models significantly outperformed the original work. These results underscore the competitiveness of our models, especially when enriched with the additional features from the proposed abusive words lexicon.

Table 1. Results of hate speech detection experiment in several Indonesian hate speech data with and without lexicon on Ibrohim and Budi [17] dataset

Model	Basic models			Basic models + Indonesian lexicon		
	macro- <i>F</i>	Acc	AUC	macro- <i>F</i>	Acc	AUC
Original work [17]	-	0.735	-	-	0.735	-
Linear SVM	0.825	0.829	0.824	0.823	0.829	0.822
DT	0.783	0.790	0.782	0.789	0.795	0.788
LR	0.842	0.848	0.839	0.839	0.845	0.836
RF	0.833	0.842	0.828	0.842	0.849	0.837
LSTM	0.834	0.843	0.828	0.837	0.844	0.831
GRU	0.837	0.845	0.831	0.840	0.847	0.835

Table 2. Results of hate speech detection experiment in several Indonesian hate speech data with and without lexicon on Pratiwi *et al.* [35] dataset

Model	Basic model			Basic models + Indonesian lexicon		
	macro- <i>F</i>	Acc	AUC	macro- <i>F</i>	Acc	AUC
Original work [35]	0.657	-	-	0.657	-	-
Linear SVM	0.732	0.733	0.736	0.731	0.733	0.732
DT	0.633	0.634	0.637	0.663	0.663	0.671
LR	0.750	0.750	0.753	0.737	0.738	0.739
RF	0.698	0.698	0.704	0.732	0.733	0.735
LSTM	0.458	0.506	0.538	0.642	0.651	0.672
GRU	0.462	0.523	0.559	0.672	0.674	0.673

Table 3. Results of hate speech detection experiment in several Indonesian hate speech data with and without lexicon on Asti *et al.* [34] Javanese dataset

Model	Basic model			Basic models + Javanese lexicon		
	macro- <i>F</i>	Acc	AUC	macro- <i>F</i>	Acc	AUC
Original work [34]	0.663	-	-	0.663	-	-
Linear SVM	0.756	0.824	0.748	0.757	0.824	0.748
DT	0.701	0.778	0.700	0.711	0.784	0.711
LR	0.765	0.841	0.743	0.766	0.841	0.744
RF	0.711	0.813	0.688	0.726	0.822	0.701
LSTM	0.751	0.819	0.744	0.760	0.82	0.761
GRU	0.678	0.711	0.737	0.734	0.788	0.753

Table 4. Results of hate speech detection experiment in several Indonesian hate speech data with and without lexicon on Asti *et al.* [34] Sundanese dataset

Model	Basic model			Basic models + Sundanese lexicon		
	macro- <i>F</i>	Acc	AUC	macro- <i>F</i>	Acc	AUC
Original work [34]	0.657	-	-	0.657	-	-
Linear SVM	0.723	0.739	0.720	0.723	0.739	0.720
DT	0.685	0.699	0.685	0.690	0.706	0.688
LR	0.732	0.748	0.728	0.733	0.749	0.729
RF	0.721	0.743	0.716	0.725	0.745	0.719
LSTM	0.729	0.741	0.729	0.737	0.749	0.735
GRU	0.726	0.748	0.720	0.734	0.747	0.732

Table 5. Results of hate speech detection experiment in several Indonesian hate speech data with and without lexicon on Desrul and Romadhony [36] dataset

Model	Basic model			Basic models + Indonesian Lexicon		
	macro- F	Acc	AUC	macro- F	Acc	AUC
Linear SVM	0.674	0.916	0.642	0.675	0.914	0.646
DT	0.666	0.911	0.639	0.679	0.911	0.655
LR	0.593	0.922	0.564	0.636	0.926	0.594
RF	0.649	0.926	0.604	0.640	0.925	0.598
LSTM	0.465	0.869	0.500	0.465	0.869	0.500
GRU	0.538	0.875	0.537	0.539	0.877	0.538

Our research provides a robust analysis of the pivotal role that external knowledge from an abusive word lexicon plays in enhancing hate speech detection within low-resource Indonesian languages. The observed improvements in model performance emphasize the need for a thoughtful choice of machine learning models, taking into account the unique attributes and size of the dataset. Whether opting for traditional or deep learning models, our findings suggest that the impact of lexicon-derived features can be significant, influencing the overall effectiveness of the detection system.

Furthermore, our study underscores the necessity of addressing dataset-specific challenges, notably annotation schemes, during the development of hate speech detection models. This nuanced understanding of dataset intricacies is crucial for devising more precise, adaptable, and culturally sensitive hate speech detection systems. By acknowledging and navigating these challenges, our research not only advances the understanding of hate speech detection in low-resource languages but also lays the groundwork for future endeavors aimed at developing more context-aware and efficient solutions for diverse linguistic and cultural contexts.

5. CONCLUSION AND FUTURE WORK

In conclusion, this research introduces novel lexicons of abusive words meticulously curated to encompass Indonesian, Javanese, and Sundanese languages. The primary objective was to gauge the efficacy of this lexicon by integrating its supplementary knowledge into machine learning models for hate speech detection in Indonesian text. Through rigorous evaluations across diverse datasets, our experiments consistently demonstrated a substantial enhancement in hate speech detection performance. The integration of lexicon-derived features significantly improved the contextual understanding of hate speech, showcasing its positive impact within the low-resource Indonesian language landscape. These compelling findings underscore the potential of external knowledge, embodied in the lexicon, to fortify machine learning models. Our study contributes not only by showcasing the effectiveness of this approach but also by unveiling novel insights into mitigating challenges within low-resource linguistic contexts. As we look forward, this research sets the stage for future endeavors to refine and expand lexicon-based approaches.

Nonetheless, while our study has demonstrated the notable efficacy of the proposed abusive word lexicon, it also unveils several compelling avenues for future research and development in this dynamic domain. One promising direction involves the continual expansion of the lexicon's vocabulary. By regularly updating and augmenting the lexicon with new words and phrases relevant to evolving online discourse, researchers can ensure its sustained relevance and applicability. This proactive approach will not only keep pace with emerging forms of abusive language but also contribute to a more comprehensive and adaptable tool for hate speech detection. Furthermore, an exciting prospect for advancing the field lies in the exploration of integrating lexicon information into more intricate models, particularly those based on transformer architectures. Transformers have demonstrated unparalleled capabilities in capturing complex linguistic patterns and contextual nuances. Integrating the lexicon into such advanced models holds the potential to unlock a deeper understanding of hate speech dynamics, especially in the context of low-resource languages. These proposed directions not only align with the evolving landscape of online communication but also address the nuanced challenges inherent in hate speech detection. These endeavors promise to contribute not only to the scientific understanding of hate speech but also to the practical development of robust solutions that align with the ever-changing nature of online discourse.

ACKNOWLEDGEMENT

This work has been funded by the Indonesian Ministry of Research and Higher Education under Grant Number 182/E5/PG.02.00.PL/2023 with title “*Pendeteksian Ujaran Kebencian untuk Bahasa Code-Mixed pada Media Sosial Berbahasa Indonesia*”.




REFERENCES

- [1] M. S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing,” *Neuro-computing*, vol. 546, p. 126232, Aug. 2023, doi: 10.1016/j.neucom.2023.126232.
- [2] G. L. De la Peña Sarracén and P. Rosso, “Systematic keyword and bias analyses in hate speech detection,” *Information Processing and Management*, vol. 60, no. 5, p. 103433, 2023, doi: 10.1016/j.ipm.2023.103433.
- [3] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, and S. H. Malik, “Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques,” *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100120, 2022, doi: 10.1016/j.ijmei.2022.100120.
- [4] R. M. O. Cruz, W. V. de Sousa, and G. D. C. Cavalcanti, “Selecting and combining complementary feature representations and classifiers for hate speech detection,” *Online Social Networks and Media*, vol. 28, p. 100194, 2022, doi: 10.1016/j.osnem.2021.100194.
- [5] E. W. Pamungkas, V. Basile, and V. Patti, “A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection,” *Information Processing and Management*, vol. 58, no. 4, p. 102544, 2021, doi: 10.1016/j.ipm.2021.102544.
- [6] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, “Deep learning based fusion approach for hate speech detection,” *IEEE Access*, vol. 8, pp. 128923–128929, 2020, doi: 10.1109/ACCESS.2020.3009244.
- [7] L. Mookdarsanit and P. Mookdarsanit, “Combating the hate speech in Thai textual memes,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 21, no. 3, pp. 1493–1502, 2021, doi: 10.11591/ijeecs.v21.i3.pp1493-1502.
- [8] M. S. Ahmed, S. M. Maher, and M. E. Khudhur, “Arabic cyberbullying detecting using ensemble deep learning technique,” *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 32, no. 2, p. 1031, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp1031-1041.
- [9] M. S. Mohamed, H. Elzayady, K. M. Badran, and G. I. Salama, “A hybrid approach based on personality traits for hate speech detection in arabic social media,” *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 4, pp. 6381–6390, 2023, doi: 10.3233/JIFS-231151.
- [10] S. G. Cahyani, A. B. Wahyudi, Markhamah, and A. Sabardila, “Code mixing on news accounts catch me up! on twitter in news text learning,” in *International Conference on Learning and Advanced Education (ICOLAE 2022)*, 2023, pp. 2024–2039, doi: 10.2991/978-2-38476-086-2.162.
- [11] I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, “Label modification and bootstrapping for zero-shot cross-lingual hate speech detection,” *Language Resources and Evaluation*, pp. 1–32, 2023, doi: 10.1007/s10579-023-09637-4.
- [12] H. Madhu, S. Satapara, S. Modha, T. Mandl, and P. Majumder, “Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments,” *Expert Systems with Applications*, vol. 215, p. 119342, 2023, doi: 10.1016/j.eswa.2022.119342.
- [13] B. R. Chakravarthi et al., “Detecting abusive comments at a fine-grained level in a low-resource language,” *Natural Language Processing Journal*, vol. 3, p. 100006, 2023, doi: 10.1016/j.nlp.2023.100006.
- [14] Y. Wirawanda and T. O. Wibowo, “TWITTER: Expressing hate speech behind tweeting,” *Profetik: Jurnal Komunikasi*, vol. 11, no. 1, p. 5, 2018, doi: 10.14421/pjk.v11i1.1378.
- [15] E. Fauziati, S. Suharyanto, A. S. Syahrullah, W. A. Pradana, and I. Nurcholis, “Hate language produced by Indonesian figures in social media: from philosophical perspectives,” *Wisdom*, vol. 3, no. 2, pp. 32–47, 2022, doi: 10.24234/wisdom.v3i2.856.
- [16] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: a dataset and preliminary study,” in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2017, vol. 2018-Janua, pp. 233–238, doi: 10.1109/ICACSIS.2017.8355039.
- [17] M. O. Ibrohim and I. Budi, “Multi-label hate speech and abusive language detection in Indonesian Twitter,” in *Proceedings of the Third Workshop on Abusive Language Online*, Aug. 2019, pp. 46–57, doi: 10.18653/v1/w19-3506.
- [18] A. R. Isnain, A. Sihabuddin, and Y. Suyanto, “Bidirectional long short term memory method and Word2vec extraction approach for hate speech detection,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 14, no. 2, p. 169, 2020, doi: 10.22146/ijccs.51743.
- [19] M. A. Ibrahim, N. T. M. Sagala, S. Arifin, R. Nariswari, N. P. Murnaka, and P. W. Prasetyo, “Separating hate speech from abusive language on Indonesian Twitter,” in *2022 International Conference on Data Science and Its Applications, ICoDSA 2022*, 2022, pp. 187–191, doi: 10.1109/ICoDSA55874.2022.9862850.
- [20] N. Aulia and I. Budi, “Hate speech detection on Indonesian long text documents using machine learning approach,” in *ACM International Conference Proceeding Series*, 2019, pp. 164–169, doi: 10.1145/3330482.3330491.
- [21] I. G. M. Putra and D. Nurjanah, “Hate speech detection in Indonesian language instagram,” in *2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, 2020, pp. 413–420, doi: 10.1109/ICACSIS51025.2020.9263084.
- [22] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, “Synonym based feature expansion for Indonesian hate speech detection,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1105–1112, 2023, doi: 10.11591/ijece.v13i1.pp1105-1112.
- [23] M. Hayaty, A. D. Laksito, and S. Adi, “Hate speech detection on Indonesian text using word embedding method-global vector,” *IAES International Journal of Artificial Intelligence (IJAI)*, vol. 12, no. 3, pp. 1928–1937, 2023.
- [24] T. I. Sari, Z. N. Ardilla, N. Hayatin, and R. Maskat, “Abusive comment identification on Indonesian social media data using hybrid deep learning,” *IAES International Journal of Artificial Intelligence (IJAI)*, vol. 11, no. 3, pp. 895–904, 2022, doi: 10.11591/ijai.v11.i3.pp895-904.




- [25] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, "Emotionally informed hate speech detection: a multi-target perspective," *Cognitive Computation*, vol. 14, no. 1, pp. 322–352, 2022, doi: 10.1007/s12559-021-09862-5.
- [26] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, "SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018, pp. 1795–1802, doi: 10.1609/aaai.v32i1.11559.
- [27] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay, "Enhanced senticnet with affective labels for concept-based opinion mining," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 2–9, 2013, doi: 10.1109/MIS.2013.40.
- [28] E. Bassignana, V. Basile, and V. Patti, "Hurtlex: a multilingual lexicon of words to hurt," in *CEUR Workshop Proceedings*, vol. 2253, Accademia University Press, 2018, pp. 51–56.
- [29] E. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: a hybrid approach with deep learning and a multilingual lexicon," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, 2019, pp. 363–370, doi: 10.18653/v1/p19-2051.
- [30] A. Koufakou, E. W. Pamungkas, V. Basile, and V. Patti, "HurtBERT: incorporating lexical features with BERT for the detection of abusive language," in *Proceedings of the fourth workshop on online abuse and harms*, 2020, pp. 34–43, doi: 10.18653/v1/2020.alw-1.5.
- [31] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words? a feature-based approach," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018, vol. 1, pp. 1046–1056, doi: 10.18653/v1/n18-1095.
- [32] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012, doi: 10.1016/j.artint.2012.07.001.
- [33] M. Hayaty, S. Adi, and A. D. Hartanto, "Lexicon-based Indonesian local language abusive words dictionary to detect hate speech in social media," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 1, p. 9, 2020, doi: 10.20473/jisebi.6.1.9-17.
- [34] A. D. Asti, I. Budi, and M. O. Ibrohim, "Multi-label classification for hate speech and abusive language in Indonesian-local languages," in *2021 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2021*, 2021, pp. 1–6, doi: 10.1109/ICACSIS53237.2021.9631316.
- [35] N. I. Pratiwi, I. Budi, and I. Alfina, "Hate speech detection on Indonesian Instagram comments using FastText approach," in *2018 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2018*, 2018, pp. 447–450, doi: 10.1109/ICACSIS.2018.8618182.
- [36] D. R. K. Desrul and A. Romadhony, "Abusive language detection on Indonesian online news comments," in *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*, 2019, pp. 320–325, doi: 10.1109/ISRITI48646.2019.9034620.
- [37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [38] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
- [39] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews," *Procedia Computer Science*, vol. 179, pp. 728–735, 2021, doi: 10.1016/j.procs.2021.01.061.
- [40] H. Imaduddin, L. A. Kusumaningtiyas, and F. Y. A'la, "Application of LSTM and GloVe word embedding for hate speech detection in Indonesian Twitter data," *Ingenierie des Systemes d'Information*, vol. 28, no. 4, pp. 1107–1112, 2023, doi: 10.18280/isi.280430.

BIOGRAPHIES OF AUTHORS






Endang Wahyu Pamungkas    is Assistant Professor at Informatics Department, Universitas Muhammadiyah Surakarta, Indonesia. He Holds a Ph.D. degree in Computer Engineering with specialization in natural language processing (NLP). He received his Bachelor and Master degree in Informatics Engineering from Institut Teknologi Sepuluh Nopember, Indonesia in 2014 and 2015 respectively. He obtained his Ph.D. in Computer Science in 2021 from University of Turin. He has authored or coauthored 23 Scopus indexed articles, with 11 H-index and more than 400 citations. His research areas are natural language processing, sentiment analysis, social media analysis, hate speech detection, and machine learning. He is the head of social informatics research center in Universitas Muhammadiyah Surakarta. He can be contacted at email: ewp123@ums.ac.id.






Dian Purworini    is a lecturer in the Communication Science Department, Faculty of Communication and Informatics, University of Muhammadiyah Surakarta. She completed her bachelor's degree at Sebelas Maret University Surakarta, master's degree at Gadjah Mada University, and received a doctorate in Communication Science from Padjadjaran University. Her research interests are crisis communication, integrated marketing communication, and social media analysis. She has had publications in both reputable international journals and national journals. In addition to pursuing research, the author is an activist in the Social Informatics Studies Center and an editor and reviewer for various journals. Previously, she served as the Head of Quality Assurance in the Communication Science Department, Vice Dean II, and Head of the Communication Science Department. She can be contacted at email: dian.purworini@ums.ac.id.



Divi Galih Prasetyo Putri    is assistant professor of Software Engineering Department in Faculty of Electronic and Informatics Engineering, Universitas Gadjah Mada. She is the Head of Software Engineering Department in Faculty of Electronic and Informatics Engineering. She has a Ph.D. in Informatics from the University of Milano Bicocca. Her main research interests include several topics in information retrieval, sentiment analysis, data science, and software engineering. She can be contacted at email: divi.galih@ugm.ac.id.



Sohail Akhtar    is currently working as an Assistant Professor at Computer Science department, Bahria University Islamabad, Pakistan. He completed his Ph.D. in Computer Science in 2021 from University of Turin, Italy. Before that, he obtained his bachelor degree from Al-lama Iqbal Open University Pakistan in 2006 and his MS degree from RWTH Aachen University Germany in 2012. His research interests include NLP, sentiment analysis, opinion polarization, abusive language and hate speech detection, and political debates. He can be contacted at email: sakhtar.buic@bahria.edu.pk.