Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities

Endang Wahyu Pamungkas[1], Divi Galih Prasetyo Putri[2], Azizah Fatmawati[3]
Informatics Engineering Department, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia[1,3]
Software Engineering Department, Vocational School, Universitas Gadjah Mada, Yogyakarta, Indonesia[2]
Social Informatics Research Center, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia[1]

Abstract—**This study aims to provide an overview of the current research on detecting abusive language in Indonesian social media. The study examines existing datasets, methods, and challenges and opportunities in this field. The research found that most existing datasets for detecting abusive language were collected from social media platforms such as Twitter, Facebook, and Instagram, with Twitter being the most commonly used source. The study also found that hate speech is the most researched type of abusive language. Various models, including traditional machine learning and deep learning approaches, have been implemented for this task, with deep learning models showing more competitive results. However, the use of transformer-based models is less popular in Indonesian hate speech studies. The study also emphasizes the importance of exploring more diverse phenomena, such as islamophobia and political hate speech. Additionally, the study suggests crowdsourcing as a potential solution for the annotation approach for labeling datasets. Furthermore, it encourages researchers to consider code-mixing issues in abusive language datasets in Indonesia, as it could improve the overall model performance for detecting abusive language in Indonesian data. The study also suggests that the lack of effective regulations and the anonymity afforded to users on most social networking sites, as well as the increasing number of Twitter users in Indonesia, have contributed to the rising prevalence of hate speech in Indonesian social media. The study also notes the importance of considering code-mixed language, out-of-vocabulary words, grammatical errors, and limited context when working with social media data.**

Keywords*—*Abusive language; hate speech detection; machine learning; social media

I. Introduction

In this digital era, social media has become an important aspect of everyday life. Not only is it a source of information, but it is also a medium of entertainment, allowing people to share content and express their feelings about anything at any time. However, social media can also be a double-edged sword. On one hand, it can provide a medium for constructive and positive communication among its users. On the other hand, the freedom of expression afforded to social media users can also create serious problems, such as the increasing prevalence of hate speech on social media. This phenomenon is often attributed to the lack of effective regulations and the anonymity afforded to users on most social networking sites [1]. These characteristics make social media the perfect medium for individual abusive users or even hate groups to spread and reinforce their views. In fact, social media platforms even offer opportunities for violent actors to propagate their acts, potentially reaching a wider audience when their posts go *viral* [2]. Twitter is a popular social networking platform that provides convenient access to its users for online social interactions. The number of Twitter users has been steadily increasing, from around 100 million users in 2017 to almost 240 million in 2022. Previous studies have shown that hate speech is also a prominent challenge in the Twittersphere. Pamungkas et al. [3] conducted a study on hate speech towards women in Twitter in multiple languages, including Italian, Spanish, and English. Lingiardi et al. [4] has also explored other forms of hate speech targeted at specific groups on Twitter.

Automatically detecting hate speech from social media text is a challenging task. Several studies have been proposed to address hate speech in social media, mainly focusing on implementing machine learning models to automatically predict whether an utterance is hate speech or not. However, working with social media data is a very challenging task. Social media data often contains valuable knowledge for information extraction tasks, but it is usually very noisy and full of informal language [5]. According to the study of Baldwin et al. [5], there are several properties of social media data, including: i) the presence of code-mixed language; ii) the presence of out-of-vocabulary words; and iii) grammatical errors. Social media data also usually has very limited context, which is an important issue for abusive language detection tasks because it is difficult to classify a text as abusive or not without context. Other important clues for abusive detection tasks, such as facial expressions, gestures, and voice tones (which are recognized in face-to-face communication), are also absent in social media data. However, social media content has some signals that can be exploited to partially resolve the context of such texts, including emojis, emoticons, hashtags, URLs, and mentions. Some studies have also found that there are several issues that contribute to the difficulty of detecting hate speech in social media automatically, including the use of swear words [6], multidomain issues [7], [8], and multilingual issues [8], [7].

Similarly, hate speech phenomena also occur in Indonesian social media. According to Statista[1], the number of Twitter users in Indonesia has reached almost 240 million, ranking fifth among all countries in the world. Hate speech in Indonesia has been regulated by the government since 2008, as stated in the Law of Information and Electronic Transaction (UU ITE). The Kepolisian Republik Indonesia (Indonesian Police

This work has been funded by the internal funding of Universitas Muhammadiyah Surakarta under Grant Number 110.28/A.3-III/LRI/VI/2022.

[1] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

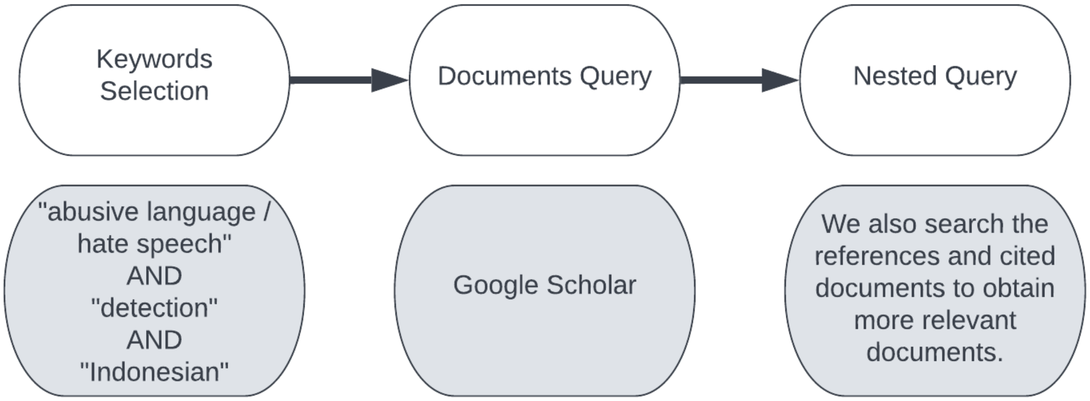
Keywords Selection
Documents Query
Nested Query
"abusive language / hate speech" AND "detection" AND "Indonesian"
Google Scholar
We also search the references and cited documents to obtain more relevant documents.

Fig. 1. Documents collection methodology.

Department) has also issued further regulations, as hate speech has the potential to have dangerous effects, not only for the victims of hate speech but also for society as a whole. Interestingly, most instances of hate speech on Indonesian social media are triggered by political events, such as elections. Several studies have also been conducted to study the hate speech phenomena in Indonesian social media [9], [10]. Most studies have focused on the automatic detection of hate speech utterances from social media data. The study by Alfina et al. [11] was one of the early studies of hate speech detection in Indonesian social media, specifically on the Twitter platform. This work proposed a novel dataset gathered from Twitter and manually annotated with two labels: hate speech and not. Another study by Ibrohim and Budi [12] proposed a more fine-grained hate speech dataset, which not only contains a binary class (hate speech vs. not), but also is annotated based on several categories, including the hate speech target, category, and level of hatefulness. More recent studies on hate speech detection in Indonesia have focused on adopting more recent technologies, such as neural-based and transform-based models [13], [14].

In this paper, we summarize the studies on hate speech detection, specifically on Indonesian social media. In this paper, we provide an overview of research conducted in this area, giving a comprehensive view of the state-of-the-art and datasets centered on this area. Our main objective is to draw a conclusion on the state-of-the-art and to provide several possible opportunities for future work based on existing open problems. After the introduction, we discuss the existing studies on hate speech detection in Indonesian social media, focusing on the approaches adopted and the available language resources in Section II. An analysis of challenges and opportunities for this particular task in future work is discussed in Section III. Finally, Section IV presents conclusive remarks for this survey.

II. Literature Review

Similar to other languages, hate speech is becoming a relevant issue in Indonesian social media. Despite being regulated by the national constitution, Indonesian social media users still use abusive language to communicate and even attack other users, often because they can hide their identities using anonymous accounts. Several studies have been conducted to address hate speech in Indonesian. Some have proposed novel corpora containing manually annotated data gathered from social media platforms such as Twitter, Instagram, and YouTube. Others have focused on developing machine learning models to automatically classify given utterances as either abusive or not. A few studies have done both, proposing a novel hate speech dataset and building a machine learning model based on that dataset. In this section, we review hate speech studies in Indonesian social media, focusing on two main aspects: (i) what datasets are available for abusive language detection in Indonesia and (ii) what has been done so far in Indonesian abusive language detection studies. We collected relevant documents using Google Scholar by searching for the keywords 'hate speech detection in Indonesian' and 'abusive language detection in Indonesian's, limited to the first five pages of results for each keyword and sorted by relevance, without a time filter. We also checked the cited documents and references on the first five pages of each search to find more relevant publications. Fig. 1 summarizes our approach to collecting relevant documents for our study.

A. What Datasets are Available for Abusive Language Detection in Indonesia?

In this subsection, we collect information about the available datasets for abusive language studies in Indonesia. Table I summarizes the information about the available datasets for hate speech detection studies specifically in Indonesian. We gathered this information from previous studies on hate speech detection in Indonesian, using the approaches outlined in Fig. 1. We found that the two most frequently used datasets in previous work are those from Alfina et. al. [11] and Ibrohim et. al. [15]. However, these datasets are still less commonly used compared to hate speech datasets in languages with more resources, such as English, Italian, and Spanish. This may be due to the lack of a hate speech detection shared task in Indonesia, which usually attracts more researchers

TABLE I. SUMMARIZATION OF AVAILABLE ABUSIVE LANGUAGE DATASET IN INDONESIAN

Topical Focus	Sources	Annotation	Entries	Available	Ref
Hate Speech	Twitter	Expert Manual	1,100	Yes	[11]
Hate Speech	Twitter	Expert Manual	13,169	Yes	[12]
Abusiveness	Twitter	Expert Manual	2,016	Yes	[15]
Abusiveness	News Comments	Expert Manual	3,184	Yes	[16]
Hate Speech	News Comments	Expert Manual	3,614	No	[16]
Hate Speech	Twitter	Expert Manual	4,002	No	[17]
Hate Speech	Instragram	Expert Manual	1,067	No	[18]
Hate Speech	Instragram	Expert Manual	13,194	No	[19]
Hate Speech	Instragram	Expert Manual	572	Yes	[20]
Hate Speech	Instragram	Expert Manual	1,012	No	[21]
Hate Speech and Cyberbullying	Twitter	Automatic	83,752	No	[22]
Hate Speech	Facebook	Expert Manual	1,276	No	[23]
Hate Speech	Twitter	Expert Manual	35,623	Yes	[24]
Hate Speech	Twitter	Expert Manual	1,477	Yes	[25]
Hate Speech	Multiple Social Media Sources	Expert Manual	2,273	No	[26]
Hate Speech	Multiple Social Media Sources	Expert Manual	1,400	No	[27]
Abusive Language and Hate Speech	Twitter	Expert Manual	5,656	Yes	[28]
Hate Speech	Twitter	Expert Manual	20,601	No	[29]

to use available datasets for developing the best systems. In this section, we will discuss the available datasets based on their topical focus, sources, annotation approach, number of instances, and availability.

- **Topical Focuses** : As mentioned in a previous study by Pamungkas et al. [8], the topical focus of a dataset can be described as the specific abusive phenomena addressed, as well as the targets of the abusive behavior. We also agree that a hate speech dataset may cover more than one abusive phenomena. Compared to the results obtained by Pamungkas et al. [8], most abusive language datasets in Indonesia only focus on two topical focuses: abusiveness and hate speech, which are the most general terms used in abusive language studies. Only one study by Febriana et al. [22] includes the term "cyberbullying" to describe their dealt abusive phenomena. Based on these results, we argue that there are still many specific abusive phenomena that need to be addressed in Indonesian abusive language studies, such as sexism, xenophobia, offensiveness, and Islamophobia.
- **Sources** : The source of a dataset refers to the media platforms from which the data was gathered. The different characteristics of each platform can also be variables that influence the treatment and difficulty of the hate speech detection task. According to our results presented in Table I, most abusive language datasets in Indonesian were gathered from Twitter. This may be due to the convenience of scraping tweet samples using the Twitter API, and because Twitter has less strict rules regarding data sharing for research purposes compared to other platforms. This result is consistent with a survey conducted by Pamungkas et al. [8]. Additionally, we also observed that some research used Instagram posts and comments on news posts to study abusive phenomena.
- **Annotation Approach and Scheme** : Based on our manual inspection of previous studies, we found that almost all of the proposed datasets were annotated by experts. This result differs from other studies of abusive language in other languages, where crowdsourcing is also a popular method for annotating datasets. We also observed that most proposed abusive language datasets in Indonesia use binary labels, including an "abusive" class and a "not abusive" class. However, we also found studies that propose a finer-grained annotation schema, such as the one implemented by [12], [28], [24]
- **Availability**: As presented in Table I, more than half of the datasets used for abusive language detection studies were not publicly available [2]. We can observe that most of the publicly available datasets were gathered from Twitter. Meanwhile, datasets sourced from other social media platforms such as Facebook and Instagram are mostly not shared publicly. This finding is also consistent with the survey results obtained by [8], where the availability of the datasets can be influenced by the regulation of the social media platforms related to data sharing policies.

B. What has been Done so Far in Indonesian Abusive Language Detection Study?

In this subsection, we review the approaches adopted by previous studies to detect abusive language in Indonesian social media. We used a similar approach as presented in Fig. 1 to collect the available studies. We collected any publications found on Google Scholar using the defined keywords, "abusive language detection Indonesia" and "hate speech detection Indonesia". We limited our query to the first five pages for each keyword and sorted results based on relevance, without a time filter. Furthermore, we also checked each document's cited documents and references on the first five pages to find more relevant publications. Table II summarizes the available works in abusive language detection, specifically in Indonesian social media. We carefully reviewed each document to obtain the key information of each work. In this part, we focus on

[2] the link of cannot be found in the article.

TABLE II. SUMMARIZATION OF APPROACHES ADDOPTED FOR HATE SPEECH DETECTION IN INDONESIAN

Model	Approach	Ref
Traditional Models	Using classical machine learning models such as SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, K-nearest Neighbours, Maximum Entropy and etc. coupled with several features including lexical and other structural features.	[11], [30], [17], [23], [12], [28], [27], [21], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]
Unsupervised Approach	Using data mining technique such as clustering, classification, and association, without training process to detect hate speech instance. This approach is very beneficial when the training data is limited.	[41], [42], [43]
Neural-based Models	Using neural-based models either RNN-based model variants such as LSTM, GRU, and etc or CNN-based models coupled with language representation either using available pretrained models or self-training based on the available training data.	[26], [19], [13], [44], [45], [46], [47], [48], [49], [50], [51], [52]
Transformer-based Models	Using the recent transformer based architecture such as BERT, RoBERTa, XLM, and etc. Based on the previous studies in NLP area, these models usually provide the robust performance across different NLP tasks.	[53], [31], [14]

reviewing the adopted approach of each work to deal with the abusive language detection task. In particular, we focus on two main discussions: variants of the models and implemented approaches. Following, we provide a deeper elaboration to compare the previous work in Indonesian abusive language studies, to gain insights for further development.

- **Model Variant**: A wide variety of classification models have been adopted for the abusive language detection task in Indonesian. Table II summarizes the published studies in this topic. Based on the results, we divided the proposed models into four different variants: traditional models, unsupervised models, neural-based models, and transformer-based models. We can observe that most previous works employed traditional models to deal with abusive language detection in Indonesia. Additionally, we also found a few studies that adopt an unsupervised approach, which do not require labeled data to detect abusive language. This is an interesting finding, as unsupervised models are not popularly used for detecting abusive language in more resource-rich languages, as observed by Pamungkas et al. [8]. Similar to traditional-based models, neural-based models are also popular for detecting abusive language. This is in line with the availability of Indonesian language models that have been proposed by several recent works. Lastly, we notice that the use of transformer-based models is still not yet explored in Indonesian abusive language studies. Unlike Indonesian language models, studies focused on developing transformer models for the Indonesian language are also limited. Most of the abusive language studies in Indonesia that exploit transformer-based models are utilizing multilingual transformers.

- **Classification Models**: A wide variety of classification models were used in this task. Starting with traditional classifiers, several models such as SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, KNN, and Maximum Entropy have been used for this classification task. These models were the most popular approach for detecting abusive language, specifically in Indonesian data. This may be due to the limited availability of resources in Indonesian, such as language models or labeled datasets. For the unsupervised approach, a few studies have proposed using lexicon-based and straightforward string matching approaches to detect abusive instances. Despite lexicon-based approaches being unpopular in common text classification tasks, this approach is still reliable when annotated data is limited. In line with the trend in other natural language processing tasks, the use of neural-based models is also gaining more attention from NLP researchers in Indonesia. Some models such as LSTM, GRU, and CNN have been widely used to detect abusive language in Indonesian, either using pre-trained language representations or without pre-training models. Lastly, the recent transformer-based technology is also starting to be used in the Indonesian research community. This may be due to the availability of multilingual transformer models such as BERT Multilingual, Multilingual GPT, and XLM RoBERTa. Some of these models were also used by a few studies [14], [53] for detecting abusive language in Indonesian.

III. CHALLENGES AND OPPORTUNITIES

The literature review and analysis presented in previous sections provide insights into the current state of the art of abusive language studies in Indonesian. Based on these analyses, we have observed several challenges in this task, which are summarized as follows:

- **Limited Availability of Language Resources**: The adopted approach for dealing with the task of abusive language detection in Indonesian is currently limited and lags behind studies in other, more resource-rich languages. Traditional models are the most popular approach for addressing this problem in Indonesia, while in other languages, more recent transformer-based models are commonly used to achieve state-of-the-art results. We believe that this discrepancy is likely related to the limited availability of language resources, including language corpora and language models. We also note that several recent studies have proposed transformer-based models, such as IndoBERT [54] and IndoBERTweet [55], but they are still limited in comparison to the transformer technologies available for other languages.

- **Limited Exploration of Abusive Phenomena**: Based on the abusive phenomena covered in the available datasets for abusive language detection studies, we perceive that the explored abusive phenomena in Indonesian is still very limited. Studies in Indonesian have mostly focused on the detection of hate and abusive speech. Meanwhile, similar studies in other languages have been conducted with a broader coverage of abusive phenomena, which can include sexism, racism, misogyny, Islamophobia, and more. Some of these studies have also proposed finer-grained labels to capture more specific abusive phenomena, which is usually beneficial for differentiating the treatment for handling each phenomenon.

- **Low Awareness of Reproducibility Aspect**: Based on our review, we also notice that most of the published research in Indonesian abusive language studies do not make their code and datasets publicly available. This issue makes it difficult for other researchers to reproduce the results of previous works, which is important for better analysis of their own studies. Furthermore, reproducibility is an important aspect for maintaining continuity in research, specifically in the area of abusive language research.
- **Limited Approach for Annotation Procedure**: We observe that most studies used manual expert annotation procedures to label abusive language datasets. This approach is proven to be reliable for obtaining a high-quality dataset when the subjectivity of the annotation task is high. However, this approach is usually not feasible for annotating a large number of data, as the annotation task becomes more labor-intensive and time-consuming. Sometimes, alternative annotation approaches such as crowdsourcing scenarios can provide a wider perspective, with a diverse demographic of annotators who have different backgrounds and views to evaluate the abusive instances.
- **The Problem of Code-Mixed Languages**: Geographically, Indonesia consists of several regions, each with its own local languages. According to recent reports, there are 718 local languages used by different regions and tribes in Indonesia. Indonesians tend to use a mix of their own local language and Bahasa to communicate on social media platforms, such as Twitter. Related to this issue, we conducted a random check on some publicly available datasets. We found a lot of code-mixed instances on the checked datasets [28], [24], which are mostly written in a mixture of Indonesian and Javanese. As in other languages and other NLP tasks, the issue of code-mixing is still a prominent challenge that needs to be tackled.

Based on these challenges, we also point out several opportunities for future studies in this research direction, which are summarized below.

- **Building Novel Language Resources in Indonesian**: Our NLP research community should also focus on studying and developing language resources in Indonesian. These resources could include novel corpora for diverse tasks or recent language model technologies. The availability of more language resources could provide more opportunities for researchers in abusive language studies to explore more approaches to better detect abusive language in Indonesian.
- **Expanding the Study Exploration into Other Abusive Phenomena**: As mentioned in the challenges section, abusive language studies in Indonesian are still focused on a few phenomena, including hate speech and abusiveness. Based on our investigation, there are several abusive phenomena specific to Indonesia that could potentially become a focus for exploration, including islamophobia and political hate speech. There are also other more general phenomena which have been studied in other languages, such as sexism, racism, xenophobia, homophobia, and more. A broader exploration into other abusive phenomena could open more opportunities for research collaboration between NLP researchers and researchers from other communities such as the study of humanity, psychology, gender studies, and social science.
- **Exploring Other Annotation Approach to Build Abusive Langueage Datasets**: Most of the available abusive language datasets in Indonesian were built using expert annotation approaches. For example, crowdsourcing could be a worth-considering option to be implemented for annotating abusive language datasets. Because crowdsourcing approach has the advantage of bringing in a diverse set of annotators with different background identities, which can help to reduce bias in the dataset, which is also an important issue in this study. In addition, crowdsourcing can be particularly useful when the dataset is large and complex, and would be too time-consuming for a single person to finish.
- **Tackling the Problem of Code-Mixed Data**: Code-mixing is becoming a prominent challenge in various NLP tasks in recent years. This problem may be due to the current technology and platforms which have a multilingual environment. Similarly, Indonesians also tend to use a mix of their local languages and Bahasa Indonesia to communicate with others both in real life and on social media channels. Dealing with language-shift in code-mixed data is a challenging task. Specifically in abusive language studies, several transfer learning approaches could be applied in this task.

IV. Conclusion

This survey presents a summary of research on detecting abusive language in Indonesian social media. It covers existing datasets that could be used for this research, including datasets from multiple platforms, types of abusive behavior, and languages. The survey also examines the methods that have been proposed for detecting abusive language in Indonesian social media. Finally, it discusses the challenges and opportunities in this area of research and provides suggestions for future development.

This study found that most of the existing datasets for detecting abusive language were collected from social media platforms like Twitter, Facebook, and Instagram, with Twitter being the most commonly used source. This may be because it is easy to obtain samples from Twitter using its public API and because of the less strict policy from Twitter for sharing data. The study also observed that hate speech is the most researched type of abusive language, compared to other types such as abusiveness and cyberbullying.

A wide variety of models have been implemented to deal with the task of abusive language detection in Indonesia. However, most studies have exploited traditional models such as logistic regression, SVM, naive bayes, and random forest to deal with this task. Several feature representations were used to train the models, which include TF-IDF, Bag of Words, and word vectors obtained from pre-trained language representations. Overall, recent deep learning architectures have obtained more competitive results compared to other models. Furthermore, we also observed that the use of transformer-based models is less popular in Indonesian hate speech studies.

Finally, we have identified some recent challenges and opportunities for abusive language detection studies in Indonesian. We observe that the availability of more language resources in Indonesian is one of the factors that contribute to the acceleration of research development, specifically in this area. We also identify that abusive language studies should explore more diverse phenomena beyond hate speech and abusiveness topics, such as islamophobia, political hate speech, and other more general phenomena which are already widely studied in other languages such as sexism and racism. Another suggestion is related to the annotation approach for labeling abusive datasets, which mostly exploit manual expert annotation procedures. We suggest exploring crowdsourcing scenarios which could produce less bias and more comprehensive datasets. Finally, we also encourage researchers who focus in this research area to consider the code-mixing issue in current abusive language datasets in Indonesia. We believe that dealing with code-mixing issue could improve the overall model performance for detecting abusive language in Indonesian data.

Acknowledgment

This research is supported by internal funding from Universitas Muhammadiyah Surakarta.

References

[1] H. Rainie, J. Q. Anderson, and J. Albright, *The future of free speech, trolls, anonymity and fake news online*. Pew Research Center Washington, DC, 2017.

[2] B. Mathew, N. Kumar, P. Goyal, A. Mukherjee *et al.*, "Analyzing the hate and counter speech accounts on Twitter," *arXiv preprint arXiv:1812.02712*, 2018.

[3] E. W. Pamungkas, V. Basile, and V. Patti, "Misogyny detection in twitter: a multilingual and cross-domain study," *Information Processing & Management*, vol. 57, no. 6, p. 102360, 2020.

[4] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis," *Behaviour & Information Technology*, vol. 39, no. 7, pp. 711–721, 2020.

[5] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang, "How noisy social media text, how diffrnt social media sources?" in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 356–364. [Online]. Available: https://www.aclweb.org/anthology/I13-1041

[6] E. W. Pamungkas, V. Basile, and V. Patti, "Investigating the role of swear words in abusive language detection tasks," *Language Resources and Evaluation*, pp. 1–34, 2022.

[7] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4675–4684. [Online]. Available: https://aclanthology.org/D19-1474

[8] E. W. Pamungkas, V. Basile, and V. Patti, "Towards multidomain and multilingual abusive language detection: a survey," *Personal and Ubiquitous Computing*, pp. 1–27, 2021.

[9] Y. Wirawanda and T. O. Wibowo, "Twitter: expressing hate speech behind tweeting," *Profetik: Jurnal Komunikasi*, vol. 11, no. 1, pp. 5–11, 2018.

[10] 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.

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

[12] 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*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 46–57. [Online]. Available: https://www.aclweb.org/anthology/W19-3506

[13] 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, pp. 169–178, 2020.

[14] 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)*. IEEE, 2022, pp. 187–191.

[15] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in indonesian social media," *Procedia Computer Science*, vol. 135, pp. 222–229, 2018.

[16] D. R. K. Desrul and A. Romadhony, "Abusive language detection on indonesian online news comments," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2019, pp. 320–325.

[17] T. Putri, S. Sriadhi, R. Sari, R. Rahmadani, and H. Hutahaean, "A comparison of classification algorithms for hate speech detection," in *Iop conference series: Materials science and engineering*, vol. 830, no. 3. IOP Publishing, 2020, p. 032006.

[18] A. Briliani, B. Irawan, and C. Setianingsih, "Hate speech detection in indonesian language on instagram comment section using k-nearest neighbor classification method," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2019, pp. 98–104.

[19] 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)*. IEEE, 2020, pp. 413–420.

[20] 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)*. IEEE, 2018, pp. 447–450.

[21] E. Erizal, B. Irawan, and C. Setianingsih, "Hate speech detection in indonesian language on instagram comment section using maximum entropy classification method," in *2019 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2019, pp. 533–538.

[22] T. Febriana and A. Budiarto, "Twitter dataset for hate speech and cyberbullying detection in indonesian language," in *2019 International Conference on Information Management and Technology (ICIMTech)*, vol. 1. IEEE, 2019, pp. 379–382.

[23] N. Aulia and I. Budi, "Hate speech detection on indonesian long text documents using machine learning approach," in *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, 2019, pp. 164–169.

[24] 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)*. IEEE, 2021, pp. 1–6.

[25] A. Muzakir, K. Adi, and R. Kusumaningrum, "Classification of hate speech language detection on social media: Preliminary study for improvement," in *International Conference on Networking, Intelligent Systems and Security*. Springer, 2023, pp. 146–156.

[26] T. L. Sutejo and D. P. Lestari, "Indonesia hate speech detection using deep learning," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 39–43.

[27] U. A. N. Rohmawati, S. W. Sihwi, and D. E. Cahyani, "Semar: An interface for indonesian hate speech detection using machine learning," in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2018, pp. 646–651.

[28] S. D. A. Putri, M. O. Ibrohim, and I. Budi, "Abusive language and hate speech detection for javanese and sundanese languages in tweets: Dataset and preliminary study," in *2021 11th International Workshop on Computer Science and Engineering, WCSE 2021*. International Workshop on Computer Science and Engineering (WCSE), 2021, pp. 461–465.

[29] F. Anistya, E. B. Setiawan *et al.*, "Hate speech detection on twitter in indonesia with feature expansion using glove," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1044–1051, 2021.

[30] N. A. Setyadi, M. Nasrun, and C. Setianingsih, "Text analysis for hate speech detection using backpropagation neural network," in *2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*. IEEE, 2018, pp. 159–165.

[31] A. D. Sanya and L. H. Suadaa, "Handling imbalanced dataset on hate speech detection in indonesian online news comments," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2022, pp. 380–385.

[32] P. S. B. Ginting, B. Irawan, and C. Setianingsih, "Hate speech detection on twitter using multinomial logistic regression classification method," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2019, pp. 105–111.

[33] M. A. Ibrahim, S. Arifin, I. G. A. A. Yudistira, R. Nariswari, A. A. Abdillah, N. P. Murnaka, and P. W. Prasetyo, "An explainable ai model for hate speech detection on indonesian twitter," *CommIT (Communication and Information Technology) Journal*, vol. 16, no. 2, pp. 175–182, 2022.

[34] D. A. Anggoro and D. Permatasari, "Performance comparison of the kernels of support vector machine algorithm for diabetes mellitus classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023.

[35] I. M. A. Niam, B. Irawan, C. Setianingsih, and B. P. Putra, "Hate speech detection using latent semantic analysis (lsa) method based on image," in *2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*. IEEE, 2018, pp. 166–171.

[36] M. P. K. Dewi and E. B. Setiawan, "Feature expansion using word2vec for hate speech detection on indonesian twitter with classification using svm and random forest," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 2, pp. 979–988, 2022.

[37] E. Utami, A. F. Iskandar, S. Raharjo *et al.*, "Multi-label classification of indonesian hate speech detection using one-vs-all method," in *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 2021, pp. 78–82.

[38] D. Elisabeth, I. Budi, and M. O. Ibrohim, "Hate code detection in indonesian tweets using machine learning approach: A dataset and preliminary study," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2020, pp. 1–6.

[39] S. Kurniawan and I. Budi, "Indonesian tweets hate speech target classification using machine learning," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*. IEEE, 2020, pp. 1–5.

[40] M. O. Ibrohim, M. A. Setiadi, and I. Budi, "Identification of hate speech and abusive language on indonesian twitter using the word2vec, part of speech and emoji features," in *Proceedings of the International Conference on Advanced Information Science and System*, 2019, pp. 1–5.

[41] W. Darmalaksana, F. Irwansyah, H. Sugilar, D. Maylawati, W. Azis, and A. Rahman, "Logical framework for hate speech detection on religion issues in indonesia," in *IOP Conference Series: Materials Science and Engineering*, vol. 1098, no. 3. IOP Publishing, 2021, p. 032046.

[42] N. Kurniasih, L. A. Abdillah, I. K. Sudarsana, I. Yogantara, I. Astawa, R. F. Nanuru, A. Miagina, J. O. Sabarua, M. Jamil, J. Tandisalla *et al.*, "Prototype application hate speech detection website using string matching and searching algorithm," *International Journal of Engineering & Technology*, vol. 7, no. 2.5, pp. 62–64, 2018.

[43] 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, pp. 9–17, 2020.

[44] J. Patihullah and E. Winarko, "Hate speech detection for indonesia tweets using word embedding and gated recurrent unit," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 1, pp. 43–52, 2019.

[45] S. S. Syam, B. Irawan, and C. Setianingsih, "Hate speech detection on twitter using long short-term memory (lstm) method," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 2019, pp. 305–310.

[46] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, "Synonym based feature expansion for indonesian hate speech detection." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, no. 1, 2023.

[47] H. Imaduddin, S. Fauziati *et al.*, "Word embedding comparison for indonesian language sentiment analysis," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*. IEEE, 2019, pp. 426–430.

[48] A. Marpaung, R. Rismala, and H. Nurrahmi, "Hate speech detection in indonesian twitter texts using bidirectional gated recurrent unit," in *2021 13th International Conference on Knowledge and Smart Technology (KST)*. IEEE, 2021, pp. 186–190.

[49] G. B. Herwanto, A. M. Ningtyas, K. E. Nugraha, and I. N. P. Trisna, "Hate speech and abusive language classification using fasttext," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2019, pp. 69–72.

[50] D. A. N. Taradhita and I. Darma Putra, "Hate speech classification in indonesian language tweets by using convolutional neural network." *Journal of ICT Research & Applications*, vol. 14, no. 3, 2021.

[51] E. Sazany and I. Budi, "Hate speech identification in text written in indonesian with recurrent neural network," in *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*. IEEE, 2019, pp. 211–216.

[52] M. N. Ramadhan, I. Budi, A. B. Santoso, and R. R. Suryono, "Sexual violence classification as hate speech using indonesian tweet," in *2022 International Symposium on Information Technology and Digital Innovation (ISITDI)*. IEEE, 2022, pp. 114–120.

[53] 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 & Management*, vol. 58, no. 4, p. 102544, 2021.

[54] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 757–770.

[55] F. Koto, J. H. Lau, and T. Baldwin, "Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10 660–10 668.