Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 22, No. 3, July 2023, pp. 495∼504
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v22i3.2939

Hate Speech Detection for Banjarese Languages on Instagram Using Machine Learning Methods

Muhammad Alkaff, Muhammad Afrizal Miqdad, Muhammad Fachrurrazi, Muhammad Nur Abdi, Ahmad Zainul Abidin, Raisa Amalia
Universitas Lambung Mangkurat, Banjarmasin, Indonesia

Article Info

Article history:

Received May 02, 2023
Revised May 30, 2023
Accepted July 01, 2023

Keywords:

Banjarese Language
Dataset
Hate Speech Detection
Instagram
Machine Learning

ABSTRACT

Hate speech refers to verbal expression or communication that aims to provoke or discriminate against individuals. The Ministry of Communication and Information of Indonesia has encountered and dealt with 3,640 cases of hate speech transmitted through digital channels between 2018 and 2021. Particularly in South Kalimantan, hate speech in the local language, Banjarese has become increasingly prevalent in recent years. Surprisingly, there is a lack of research on using machine learning to detect hate speech in the Banjarese language, specifically on Instagram. Therefore, this study aimed to address this gap by constructing a dataset of Banjarese language hate speech and comparing various feature extraction and machine learning models to detect Banjarese language hate speech effectively. This research used several feature extraction techniques and machine learning methods to detect Banjarese language hate speech. The feature extraction methods used were Word N-Gram, Term Frequency-Inverse Document Frequency (TF-IDF), a combination of Word N-Gram and TF-IDF, Word2Vec, and Glove, while the machine learning methods used were Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. The results of this study revealed that the combination of TF-IDF for feature extraction and SVM as the model achieves exceptional performance. The average Recall, Precision, Accuracy, and F1-Score score exceeded 90%, demonstrating the model's ability to identify Banjarese hate speech accurately.

Corresponding Author:

Muhammad Alkaff, +6281953632809
Faculty of Engineering and Department of Information Technology,
Universitas Lambung Mangkurat, Banjarmasin, Indonesia,
Email: m.alkaff@ulm.ac.id

How to Cite:
M. Alkaff, M. Miqdad, M. Fachrurrazi, M. Abdi, A. Abidin, and R. Amalia, ”Hate Speech Detection for Banjarese Languages on Instagram Using Machine Learning Methods”, MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 3, pp. 495-504, Jul. 2023.

1. INTRODUCTION

Hate speech is an expression, writing, action, or performance intended to provoke violence or discrimination against someone based on the characteristics of their society; represent, such as race, ethnicity, gender, sexual orientation, religion, and other characteristics [1]. Hate speech is one of the important topics of discussion related to social media analysis. It is mainly associated with the freedom of users to share content and opinions on existing social media platforms [2]. Freedom of opinion in social media has also led to increased hate speech through social media. Hate speech containing harsh words or phrases accelerates social conflict because harsh words/phrases trigger emotions [3]. This problem affects the dynamics and interactions of online social communities. In Indonesia, the Ministry of Communication and Information Technology of the Republic of Indonesia (KOMINFO) handled 3,640 SARA-based Hate Speeches in the Digital Space from 2018 to April 26, 2021. In South Kalimantan, hate speech cases have been rampant in recent years. Quoted from several news pages in 2018, a social media account uploaded content that allegedly contained elements of hate speech that were considered insulting to a cleric from Banjar, South Kalimantan. In 2020, a State Civil Apparatus (ASN) was arrested for spreading hoax news and hate speech against the Indonesian National Police (POLRI) institution. In January 2021, when a major flood hit South Kalimantan, H Sahbirin Noor became the target of hate speech from South Kalimantan residents in his actions to deal with floods. In South Kalimantan, most of the hate speech uttered by residents of South Kalimantan uses the Banjarese language. From several social media, the most common hate speech found in it is Instagram.

Hate speech detection has become crucial in social media platforms, including Instagram. The Banjarese language is one of the languages spoken in Indonesia, and detecting hate speech in this language on Instagram is a relatively new area of research. This review aims to provide an overview of previous studies that can support and strengthen novelty's contribution to detecting the hate speech of Banjarese Language on Instagram. Previous research has extensively explored the accuracy of machine learning methods in detecting hate speech on social media. The effectiveness of these methods depends on the language and dataset used [4]. For instance, a study focused on the English language employed a dataset comprising 14,509 tweets from Twitter. The study applied the SVM Linear algorithm to classify hate speech, achieving an accuracy rate of 78%. Furthermore, a research endeavor on the Indonesian language involved a dataset of 13,169 tweets from Twitter. The study used RFDT (Random Forest Decision Tree) and LP (Linear Programming) transformation methods. Without identifying targets, categories, and levels, the classification process achieved an accuracy rate of 77.36%. In contrast, the classification with the identification of targets, categories, and levels yielded an accuracy rate of 66.12% [3]. Salim and Suhartono [5] conducted a systematic literature review of different machine-learning methods for hate speech detection. The study can be used to make an experimental approach to detecting hate speech and abusive language. Zhang et al. [2] observed that extremist violence tends to increase online hate speech, particularly on messages directly advocating violence [6]. Sinyangwe established that in the fore model, to detect hate speech and offensive language on online social media platforms, the data set must be categorized and presented in statistical form after running the model. Ghosal and Jain [7] identified the need for artificial intelligence (AI) in hate speech research. Awal [8] explored fine-tuning language models (LMs) to perform hate speech detection, and these solutions have yielded significant performance.

Li and Ning [9] researched anti-Asian hate speech detection via data-augmented semantic relation inference. Boishakhi et al. [10] Used a combined approach to detect hate speech from contents using video, audio, and speech by extracting feature images and feature values from audio and text. They used Machine learning, Deep learning, and Natural language processing to detect hate speech. In [11], the researchers used Long Short-Term Memory for hate speech and abusive language detection on Indonesian Youtube comment sections. Deshpande et al [12]. They have conducted experiments for a binary hate speech classification task in Multilingual-Train Monolingual-Test, Monolingual-Train Monolingual-Test, and Language-Family-Train Monolingual Test scenarios. Mozafari et al. [13] investigated the feasibility of applying a meta-learning approach in cross-lingual few-shot hate speech detection by leveraging two meta-learning models based on optimization-based and metric-based (MAML and Proto-MAML) methods. These findings demonstrate the varying performance of different machine learning approaches in hate speech detection, depending on the language and dataset under consideration. Therefore, the novelty of this research lies in investigating hate speech detection using machine learning techniques, specifically in the context of the Banjarese language on social media platforms. In order to address this gap in the literature, this study aims to explore existing methods and identify the most accurate approach for detecting hate speech in the Banjarese language.

The data utilized in this study comprises comments extracted from local Instagram accounts known for frequently containing hate speech. Three commonly employed models were chosen for text classification purposes: Support Vector Machine (SVM), Naïve Bayes, and Decision Tree. SVM is commonly employed as a binary classifier in natural language processing (NLP) tasks [14]. It constructs margins between classes to maximize the distance between the margins and the classes, thereby minimizing classification errors [15]. Naïve Bayes, widely recognized for its effective assumptions and ease of implementation, is extensively used for text classification [16]. Decision trees have been extensively employed in various machine learning tasks, as they possess a lucid structure that offers insights into the training data and facilitates straightforward implementation [17]. This study aims to determine the most accurate method for detecting hate speech on social media, particularly Instagram. Consequently, the findings of this research can serve as a valuable reference when selecting an appropriate machine-learning method to assess the accuracy of hate speech detection in the Banjarese language. The researchers aspire that this study will benefit other scholars, particularly those in the low-resource local language like Banjarese.

2. RESEARCH METHOD

This research aims to create a Banjarese language hate speech dataset and try several combinations of feature extraction and machine learning models to determine which combination has the best accuracy in classifying hate speech. The method used in this study can be seen in Figure 1.

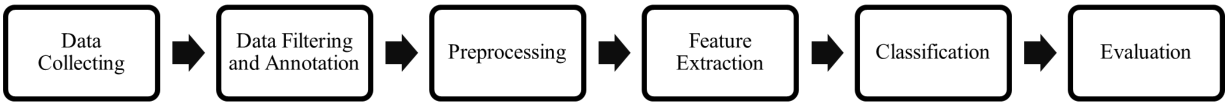

Figure 1. Research Methods

2.1. Data Collection

Because this study focuses on detecting hate speech in the Banjarese language, where previously there was no dataset, the researchers created a dataset for this study by collecting comments on local Instagram accounts where many comments were found in Banjarese. Comments are mainly collected from posts that discuss disasters, politics, or other topics that trigger hate speech.

2.2. Data Filtering and Annotation

At the data filtering stage, the researcher removed the redundant data and changed the comments in languages other than Banjarese into Banjarese in the dataset. The process of language change refers to the Banjarese language dictionary and is validated by linguists. Dataset labeling will be done manually by the researchers themselves. Labeling is done by marking each data as ”hate speech” with the number 1 or ”not hate speech” with the number 0. Before annotating the data, the researcher prepared guidelines as the rules of hate speech used in this study.

2.3. Preprocessing

Before classifying the data, it is necessary to carry out several preprocessing procedures. Case folding involves changing words in a text into uniform lowercase letters to facilitate further processing [18, 19]. Stop Word Removal, stop word is a common word that often appears in a sentence but has no meaning [18]. Removing stop words can increase the signal-to-noise ratio in unstructured text and thus increase the statistical significance of terms that may be important for a specific task [20]. Punctuation Removal, this flag - used to divide the text into sentences, paragraphs, and phrases - affects the result of any text processing approach, especially what depends on the frequency of occurrence of words and phrases because punctuation marks are often used in the text [21]. Most text and document data sets contain many unnecessary characters, such as punctuation and special characters [22]. Critical punctuation and special characters are essential for the human understanding of documents, but they can harm classification algorithms [23]. URLs Removal, URLs do not correlate with the meaning of a comment, which can reduce classification performance, and are also not used in the following process [24, 25].

2.4. Feature Extraction

Machine learning algorithms cannot understand classification rules on unprocessed text. Machine learning algorithms need numeric features to understand classification. Therefore, feature extraction is one of the main steps in text classification. This step extracts the main features from the raw text and represents the features extracted in numerical form [26]. In this research, the feature extraction used by the researcher is Word N-gram, TF-IDF, a combination of Word N-gram and TF-IDF, Word2Vec, and Glove, shown in Table 1.

Table 1. Feature Extraction (Key Concept)

Concept	Definition	References
Word N-gram	is a technique of collecting sequential word lists with sizes 1, 2, 3, N; to list all expressions of size N and calculate their frequency.	[27]
Term Frequency - Inverse Document Frequency	It is a feature representation technique representing ”word importance” to a document in the document set. It works in a combination of the frequency of word appearance in a document with no. of documents containing that word.	[28]
Word2vec	It is a technique to learn vector representation of words, which can further be used to train machine learning models.	[29]
Glove	Global log bilinear regression model that combines the advantages of the two main model families in literature: global matrix factorization and local context window method	[30]

2.5. Classification

The researcher classified the data by dividing the data into several classes, with class divisions, namely: true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). Several machine learning algorithms are applied in this research: SVM, Naïve Bayes, and Decision Tree, which detect hate speech in the Banjarese language. This algorithm is implemented using the scikit learn library [31].

2.6. Evaluation

For evaluation, the researcher applies the F1-measure and Accuracy as performance evaluation metrics in this study [31]. Accuracy is the ratio of correct predictions to the total number of samples, while F1-measure is the harmonic mean of Precision and Recall. Classifier Performance is measured by calculating true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP), which will form a confusion matrix. The confusion matrix table is shown in Table 2.

Table 2. Confusion Matrix

		Actual	
		Positive	Negative
Predict	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

True Positive (TP) is the proportion of positive instances classified correctly [32]. False Positive (FP) refers to the number of incorrectly classified hate speeches [33]. False Negative (FP) is the number of incorrect dictions that an instance is negative [34]. True Negative (TN) represents the number of negative examples if the classification result is correct [35]. Different performance metrics are used to assess the performance of the classifier that has been made. Models built in this experiment were evaluated by calculating their F1-score [36, 37]. Some performance details metrics are discussed briefly below [26]. The accuracy rate is the total number of correctly classified over the total number of samples (true positives and true negatives) [26, 38]. The formula for the accuracy rate is shown in (1). The recall is the proportion of actual positives which are predicted positive [38]. The formula for the recall rate is shown in (2). Precision is also a positive predictive value indicating the algorithm's accuracy for each model that detects hate speech [26]. The formula for the precision rate is shown in (3). F1-measure evaluates the harmonic value between recall and precision [38]. The formula for the F1-measure rate is shown in (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F1 - measure = 2 \cdot \frac{Precision \ \cdot \ Recall}{Precision \ + \ Recall} \tag{4}$$

3. RESULT AND ANALYSIS

3.1. Banjarnese Hate Speech Dataset

The Banjarese language hates speech dataset created comes from comments on local South Kalimantan Instagram accounts that speak Banjarese. The process of making this dataset goes through several stages: data collecting, data filtering and annotation, preprocessing, feature extraction, classification, and evaluation. The CSV-formatted dataset consists of 15,481 data instances, 2,039 classified as hate speech, and 13,442 as not being hate speech (See Table 3). The sample dataset and labels used in this study are shown in Table 4. Due to the data imbalance, the F1-measure metric will be used to measure accuracy. F1-measure is a composite metric considering precision and recall. Precision measures correctly predicted hate speech instances out of all predicted hate speech instances, while recall measures correctly predicted hate speech instances out of all actual hate speech instances. F1-measure provides a balanced evaluation metric, particularly for imbalanced datasets. The F1 measure enhances model performance when data imbalance is addressed appropriately [4].

Table 3. Dataset Distribution

Number of Sentences	Hate Speech	Normal Speech
15,481	2,039	13,442

Table 4. Banjarese Language Hate Speech Dataset

Text	Translation	Label
Cabut ja bungul	Just let go stupid	1
Indonesia ni gatuk pang	This is Indonesia, let's touch	0
tambuk	Dull	1
Liwar tahi	So shit	1
Handak tetawa tapi ini indonesia	I want to laugh but this is Indonesia	0
Bebanyak begal ni	More and more thugs	0
Mehadangi habar berita nang hanyar admin	Waiting for the latest news admin	0
Mirisnya hukum negara kaini	It's sad that state law like this	0
Negara Indonesia lucu tapi bungul	Indonesia is funny but stupid	1
Polisi nang bepandir bungul	Stupid talking cop	1

3.2. The Combination of Feature Extraction and Model for Detecting Banjarese Hate Speech

After the dataset is collected, the next step is to perform feature and model extraction and then compare the combination of feature and model extraction with the Recall, Precision, Accuracy, and F1-Measure metrics to find the most accurate combination of feature extraction and model in detecting hate speech in Banjarese language. The dataset created was divided into 8:2 compositions for training and testing compositions. The results of combining feature extraction and models using the dataset created can be seen in Table 5. In the accuracy metric, the combination of feature extraction and model with the highest score after being applied to the Banjarese language hate speech dataset is TF-IDF and SVM, with a score of 91%. In the recall metric, there are two feature extraction combinations, and the model with the highest score with the same number. TF-IDF and Naïve Bayes, as well as TF-IDF and SVM, are the combination of feature extraction and model that has the highest score after being applied to the Banjarese language hate speech dataset with the same score of 91%. In the Precision metric, there are differences between the two previous metrics. The combination of feature extraction and model with the highest score is TF-IDF and Naïve Bayes with a score of 91%. In the F1-Measure metric, SVM and TF-IDF are the combinations of feature extraction and model with the highest score after being applied to the Banjarese language hate speech dataset with a score of 91%.

Table 5. Performance of Algorithms

Models	Feature Extraction	Accuracy (%)	Recall (%)	Precision (%)	F1-measure (%)
SVM	N-Gram	90	90	89	90
	TF-IDF	**91**	**91**	90	**91**
	N-Gram & TF-IDF	90	90	89	90
	Word2Vec	88	88	89	88
	Glove	87	87	81	87
Naïve Bayes	**N-Gram**	**91**	**91**	90	90
	TF-IDF	90	90	**91**	89
	N-Gram & TF-IDF	90	90	90	89
	Word2Vec	78	78	82	78
	Glove	42	42	81	42
Decision Tree	N-Gram	89	89	88	89
	TF-IDF	89	89	88	89
	N-Gram & TF-IDF	87	87	86	87
	Word2Vec	78	78	81	78
	Glove	80	80	79	80

It can be seen from Table 5 that Naïve Bayes and SVM models with N-Gram and TF-IDF feature extraction dominate the highest values for F1-measure, Accuracy, Precision, and Recall metrics. However, due to unbalanced data, the accuracy metric used is F1-measure, so SVM and TF-IDF are the best model and combinations of feature extraction from this research to detect hate speech in the Banjarese language. Table 6 shows the comparison of this research with previous research. The research [**?**] conducts a comparative analysis of studies focusing on different languages, including Javanese, Sundanese, Madurese, Minangkabau, and Musi. In contrast, research [**?**] specifically compares previous research on Sundanese and Javanese languages. The novelty aspect of each study is emphasized in the corresponding column, and the outcomes of prior investigations are contrasted with the present study's findings. The results presented in reference [39] demonstrate a positive correlation between dataset size and performance improvement. The current study employs a Banjarese language dataset comprising 15,481 instances, achieving an F1-measure of 91%. These results indicate superior performance compared to previous studies conducted on other regional languages.

On the other hand, reference [**?**] focuses on comparing different algorithms and feature extraction techniques. The earlier research achieved F1-measures ranging from 80% to 82% using N-Gram feature extraction in combination with algorithms such as SVM, RFDT, and Naïve Bayes for Sundanese and Javanese languages. However, the present study surpasses these previous findings by employing TF-IDF feature extraction. By utilizing this approach in conjunction with SVM, the F1-measure for detecting Banjarese hate speech reaches 91%. The effectiveness of the TF-IDF feature extraction method stems from its ability to assign higher weights to words that offer greater information content within a specific document while considering their rarity across the entire dataset. This weighting scheme proves instrumental in capturing the discriminative power of words specific to hate speech in the Banjarese language. Furthermore, TF-IDF effectively mitigates the influence of common words that frequently appear in both hate speech and non-hate speech documents. By downplaying the significance of these common words, the feature extraction method can focus more on identifying distinctive words and phrases that serve as indicators of hate speech in the Banjarese language. Thus, the TF-IDF feature extraction method takes into account the distribution of words across the entire dataset to enhance hate speech detection capabilities.

Table 6. Comparison of Research Results

References	Novelty	Result (Previous Study)	Result (This Study)
[39]	**Comparison:** Based on research using other regional languages, such as the Javanese language with a dataset of 3449, the Sundanese language with a dataset of 2207, the Madurese language with a dataset of 2773, Minangkabau language with a dataset of 3125, and Musi language with a dataset of 2564. **Novelty:** The results show that a larger number of datasets increases the performance results obtained.	Java language Dataset 3449 F1-measure 87.5% Sundanese language Dataset 2207 F1-measure 79.5% Madurese language Dataset 2773 F1-measure 73.9% Minangkabau language Dataset 3125 F1-measure 69% Musi language Dataset 2207 F1-measure 80%	Banjarese language Dataset 15481 F1-measure 91%
[?]	**Comparison:** Results from previous research on Sundanese and Javanese using Naïve Bayes, SVM, and RFDT algorithms and N-Gram feature extraction yielded better performance. **Novelty:** The results of this study on Banjarese language using SVM, Naïve Bayes, and Decision Tree, as well as TF-IDF feature extraction, resulted in a much better F1 measure.	Comparison F1-measure: SVM+ N-Gram 82% RFDT + N-Gram 82% Naïve Bayes + N-Gram 80%	Comparison F1-measure: SVM+ TF-IDF 91% DT + TF-IDF 89% Naïve Bayes + TF-IDF 89%

4. CONCLUSION

This research uses feature extraction and model experiments to investigate hate speech detection in the Banjarese language. By analyzing a dataset of 15,481 instances, including 2,039 hate speech samples and 13,442 non-hate speech samples, the study finds that the combination of TF-IDF feature extraction and the Support Vector Machine (SVM) model achieves an average accuracy score exceeding 90% for each metric. The research contributes novel insights to the field by addressing the lack of previous studies in hate speech detection for the Banjarese language, and it offers practical implications for future research in refining detection methods and enhancing accuracy. The findings of this study have significant implications for hate speech detection in the Banjarese language. The demonstrated effectiveness of the TF-IDF feature extraction method and SVM model underscores their potential as accurate tools for distinguishing Banjarese language hate speech. The research also provides a valuable dataset for further exploration, enabling researchers to investigate alternative approaches and refine detection methods specific to the Banjarese language. Overall, this study expands knowledge in hate speech detection and offers valuable insights for future research endeavors in this area.

5. ACKNOWLEDGEMENTS

The authors want to acknowledge the research funding by the Directorate of Learning and Student Affairs (BELMAWA) under the Student Creativity Program (PKM) grant, number 2489/E2/KM.05.01/2022.

6. DECLARATIONS

AUTHOR CONTIBUTION

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

FUNDING STATEMENT

The Directorate of Learning and Student Affairs (BELMAWA) funded the research under the Student Creativity Program (PKM) grant, number 2489/E2/KM.05.01/2022.

COMPETING INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine Learning Techniques for Hate Speech Classification of Twitter Data: State-Of-The-Art, Future Challenges and Research Directions," *Computer Science Review*, vol. 38, p. 100311, 2020.

[2] G. H. Martono, A. Azhari, and K. Mustofa, "An Extended Approach of Weight Collective Influence Graph for Detection Influence Actor," *International Journal of Advances in Intelligent Informatics*, vol. 8, no. 1, pp. 1–11, mar 2022.

[3] 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*. Stroudsburg: Association for Computational Linguistics, 2019, pp. 46–57.

[4] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," *IEEE Access*, vol. 9, pp. 88 364–88 376, 2021.

[5] C. E. Rudy Salim and D. Suhartono, "A Systematic Literature Review of Different Machine Learning Methods on Hate Speech Detection," *JOIV : International Journal on Informatics Visualization*, vol. 4, no. 4, pp. 213–218, dec 2020.

[6] A. Olteanu, C. Castillo, J. Boy, and K. Varshney, "The Effect of Extremist Violence on Hateful Speech Online," in *Proceedings of the international AAAI conference on web and social media*, 2018, pp. 1–10.

[7] S. Ghosal and A. Jain, "Research Journey of Hate Content Detection From Cyberspace," 2021, pp. 200–225.

[8] M. R. Awal, R. K.-W. Lee, E. Tanwar, T. Garg, and T. Chakraborty, "Model-Agnostic Meta-Learning for Multilingual Hate Speech Detection," *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2023.

[9] J. Li and Y. Ning, "Anti-Asian Hate Speech Detection via Data Augmented Semantic Relation Inference," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2022, pp. 607–617.

[10] F. T. Boishakhi, P. C. Shill, and M. G. R. Alam, "Multi-modal Hate Speech Detection using Machine Learning," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, dec 2021, pp. 4496–4499.

[11] C. Erico, "Long Short-Term Memory Approach For Hate Speech and Abusive Language Detection on Indonesian Youtube Comment Section," Ph.D. dissertation, 2021.

[12] N. Deshpande, N. Farris, and V. Kumar, "Highly Generalizable Models for Multilingual Hate Speech Detection," *CoRR*, 2022.

[13] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning," *IEEE Access*, vol. 10, pp. 14 880–14 896, 2022.

[14] Y. Li, K. Bontcheva, and H. Cunningham, "Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction," pp. 1–25, 2006.

[15] B. AlBadani, R. Shi, and J. Dong, "A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM," *Applied System Innovation*, vol. 5, no. 1, p. 13, 2022.

[16] D. Jurafsky and J. Martin, "Naive Bayes and Sentiment Classification," *Speech and Language Processing*, p. 1024, 2019.

[17] M. Bansal, A. Goyal, and A. Choudhary, "A Comparative Analysis of K-Nearest Neighbour, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in Machine Learning," *Decision Analytics Journal*, p. 100071, 2022.

[18] U. A. N. Rohmawati, S. W. Sihwi, and D. E. Cahyani, "SEMAR: An interface for Indonesian hate speech detection using machine learning," *2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018*, no. 1, pp. 646–651, 2018.

[19] S. Cokrowibowo and N. Zulkarnaim, “Online News Analysis of Majene Public Figure Electability With NLP (Natural Language Processing),” in *IOP Conference Series: Materials Science and Engineering*, vol. 875, no. 1. IOP Publishing, 2020, p. 12092.

[20] S. Sarica and J. Luo, “Stopwords in Technical Language Processing,” *PLoS ONE*, vol. 16, no. 8 August, pp. 1–13, 2021.

[21] A. Garlapati, N. Malisetty, and G. Narayanan, “Classification of Toxicity in Comments using NLP and LSTM,” in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2022, pp. 16–21.

[22] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text Classification Algorithms: A Survey,” *Information (Switzerland)*, vol. 10, no. 4, pp. 1–68, 2019.

[23] B. Pahwa, S. Taruna, and N. Kasliwal, “Sentiment Analysis- Strategy for Text Pre-Processing,” *International Journal of Computer Applications*, vol. 180, no. 34, pp. 15–18, 2018.

[24] D. Z. Abidin, S. Nurmaini, R. F. Malik, E. Rasywir, and Y. Pratama, “A Model of Preprocessing for Social Media Data Extraction,” in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. IEEE, 2019, pp. 67–72.

[25] N. I. Pratiwi, I. Budi, and M. A. Jiwanggi, “Hate Speech Identification Using the Hate Codes for Indonesian Tweets,” in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. ICST, jul 2019, pp. 128–133.

[26] S. Abro, S. Shaikh, Z. Ali, S. Khan, G. Mujtaba, and Z. H. Khand, “Automatic Hate Speech Detection Using Machine Learning: A Comparative Study,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 484–491, 2020.

[27] A. Alrehili, “Automatic Hate Speech Detection on Social Media: A Brief Survey,” *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, vol. 2019-Novem, pp. 1–6, 2019.

[28] H. Zhou, “Research of Text Classification Based on TF-IDF and CNN-LSTM,” in *Journal of Physics: Conference Series*, vol. 2171, no. 1. IOP Publishing, 2022, p. 12021.

[29] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, “Sentiment Analysis on Online Transportation Reviews Using Word2vec Text Embedding Model Feature Extraction and Support Vector Machine (Svm) Algorithm,” in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*. IEEE, 2022, pp. 163–167.

[30] J. Pennington, R. Socher, and C. Manning, *Glove: Global Vectors for Word Representation*, jan 2014, vol. 14.

[31] N. Aulia and I. Budi, “Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach,” in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. ICST, apr 2019, pp. 164–169.

[32] E. M. Dharma, F. L. Gaol, H. Warnars, and B. Soewito, “The Accuracy Comparison Among Word2vec, Glove, and Fasttext Towards Convolution Neural Network (CNN) Text Classification,” *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 2, p. 31, 2022.

[33] S. Khan, A. Kamal, M. Fazil, M. A. Alshara, V. K. Sejwal, R. M. Alotaibi, A. R. Baig, and S. Alqahtani, “HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network,” *IEEE Access*, vol. 10, pp. 7881–7894, 2022.

[34] T. Pöyhönen, M. Hämäläinen, and K. Alnajjar, “Multilingual Persuasion Detection: Video Games as an Invaluable Data Source for NLP,” *arXiv preprint arXiv:2207.04453*, 2022.

[35] C. C. Wang, M. Y. Day, and C. L. Wu, “Political Hate Speech Detection and Lexicon Building: A Study in Taiwan,” *IEEE Access*, vol. 10, pp. 44 337–44 346, 2022.

[36] P. Jain, K. R. Srinivas, and A. Vichare, “Depression and Suicide Analysis Using Machine Learning and NLP,” in *Journal of Physics: Conference Series*, vol. 2161, no. 1. IOP Publishing, 2022, p. 12034.

[37] W. Etaiwi and G. Naymat, “The Impact of Applying Different Preprocessing Steps on Review Spam Detection,” *Procedia Computer Science*, vol. 113, pp. 273–279, 2017.

[38] A. A. Amri, A. R. Ismail, and O. A. Mohammad, “Evolutionary Deep Belief Networks With Bootstrap Sampling for Imbalanced Class Datasets,” *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 2, pp. 123–136, 2019.

[39] 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,” *2021 11th International Workshop on Computer Science and Engineering, WCSE 2021*, no. Wcse, pp. 461–465, 2021.