Separating Hate Speech from Abusive Language on Indonesian Twitter

Muhammad Amien Ibrahim
Department of Computer Science, School of Computer Science
Bina Nusantara University
Indonesia
muhammad.ibrahim1@binus.edu

Noviyanti Tri Maretta Sagala
Department of Statistics, School of Computer Science
Bina Nusantara University
Indonesia
noviyanti.sagala@binus.edu

Samsul Arifin
Department of Statistics, School of Computer Science
Bina Nusantara University
Indonesia
samsul.arifin@binus.edu

Rinda Nariswari
Department of Statistics, School of Computer Science
Bina Nusantara University
Indonesia
rinda.nariswari@binus.ac.id

Nerru Pranuta Murnaka
Department of Mathematics Education
STKIP Surya
Indonesia
nerru.pranuta@stkipsurya.ac.id

Puguh Wahyu Prasetyo
Mathematics Education Department
Universitas Ahmad Dahlan
Indonesia
puguh.prasetyo@pmat.uad.ac.id

2022 International Conference on Data Science and Its Applications (ICoDSA) | 978-1-6654-8665-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICoDSA55874.2022.9862850

***Abstract*—Social media is an effective tool for connecting with people and distributing information. However, many people often use social media to spread hate speech and abusive languages. In contrast to hate speech, abusive languages are frequently used as jokes with no purpose of offending individuals or groups, even though they may contain profanities. As a result, the distinction between hate speech and abusive language is often blurred. In many cases, individuals who spread hate speech may be prosecuted as it has legal implications. Previous research has focused on binary classification of hate speech and normal tweets. This study aims to classify hate speech, abusive language, and normal messages on Indonesian Twitter. Several machine learning models, such as logistic regression and BERT models, are utilized to accomplish text classification tasks. The model's performance is assessed using the F1-Score evaluation metric. The results show that BERT models outperform other models in terms of F1-Score, with the BERT-indobenchmark model, which was pretrained on social media text data, achieving the highest F1-Score of 85.59. This also demonstrates that pre-training the BERT model using social media data improves the classification model significantly. Developing such classification model that can distinguish between hate speech and abusive language would help individuals in preventing the spread of hate speech that has legal implications.**

***Keywords*—hate speech, abusive, twitter, separating**

I. INTRODUCTION

Any speech directed at a person or group that conveys hatred based on something about that person or group is considered hate speech. Ethnicity, religion, handicap, gender, and sexual orientation are all commonly used to justify hatred. Hate speech propagation is a dangerous practice that can lead to prejudice, societal turmoil, and even genocide. In ordinary life, hate speech is frequently accompanied by abusive language, particularly on social media [3]. Abusive language is an expression that incorporates offensive words or profanities aimed at individuals or groups. Hate speech that includes harsh words/phrases that provoke emotions frequently increases the initiation of social conflict [4]. In Indonesia, abusive phrases are mainly formed from an unpleasant situation such as mental illness, sexual deviation, physical impairment, a condition where someone lacks etiquette, and other conditions connected to unfortunate circumstances; animals with a negative trait; astral creatures that regularly interfere with human existence; a dirty and filthy environment. [5]. Due to the use of abusive words/phrases that stimulate emotions, the spread of hate speech accompanied with abusive language generally increases the prevalence of social conflict. [6]. Even though harsh language is sometimes used as a joke (not to insult someone), its use on social media can nevertheless cause conflict owing to misconceptions among users. Despite being relatively close, abusive language is not necessarily hate speech [7]. To reduce conflicts between individuals and children who are exposed to hate speech and abusive language from the social media they use, hate speech and abusive language on social media must be monitored [8]. In recent years, some researchers have investigated hate speech identification and abusive language detection in various methods. [9]. Hate speech has a distinct objective, classification, and degree while abusive language is not categorized into any specific target, group, or levels [4]. Hate speeches are intended towards a specific individual or group with a high level of animosity and fall under a variety of categories, including ethnicity, religion, race, sexual orientation, and others [10].

In the previous work, hate speech detection model has been developed by proposing a transfer learning strategy to improve the performance on publicly available benchmark English datasets with additional data including document containing racism, sexism, neither, or both [11]. Another work in [12] provides a new large-scale Brazilian Portuguese dataset with tweets labeled as harmful, non-toxic, or in different forms of toxicity. In the previous works, hate speech detection models have been developed by combining different form of categories such as racism, sexism, and other form of toxicity. Other hate speech detection model have also been developed by looking at granular level and categories of toxicity. [4] investigate abusive language and hate speech detection in tweets, including the target, categories, and degree of toxicity. The work in [13] explored a deeper level of target classification of hate speech for Indonesian tweets, finding that general hate speech and non-hate speech both fail to capture the core of hate speech. In spite of being able to utilize abusive language feature to complement hate speech detection model, [4] does not distinguish whether a document is hate speech or abusive language. This is important because hate speech and abusive language often overlap making the difference is quite unclear although as stated in [13], hate speech could lead to legal consequences. Therefore [14] develop a detection model to

distinguish hate speech and hateful, offensive and neutral instances on English language tweets. This study attempts to separate hate speech and abusive language on Indonesian twitter.

In Indonesia, according to national cybercrime agency and The National Commission on Human Rights as stated [4], hate speech is a harmful conduct that is directed at a specific target and can result in social conflict. In some cases, deliberately spreading hate speech also lead to legal implication [13]. In contrast, abusive language is often used in the context of humor although containing offensive language. This overlap makes it difficult to differentiate especially when hate speech carries legal consequences in some countries such as Indonesia. Many previous works only focused on binary classification of hate speech and non-hate speech. To fill the gap, this study develops detection models that classify hate speech, abusive language, and normal tweets. Several machine learning models are utilized to perform text classification and evaluation metrics such as F1-score is used to compare models' performance.

II. Methods

Hate speech is any speech directed towards individuals or groups that hateful expressions based on the characteristics of those individuals or group [4]. Hateful sentiments are frequently expressed based on race, religion, disability, and sexual. Hate speech that is expressed in public can lead to conflict in society. On the other hand, a verbal or written expression that contains offensive words with no aim of harming others and often used as sarcastic remarks is referred to abusive language [4]. Animals with negative characteristics, mental illness, sexual deviation, disability, other unfortunate circumstances, and illnesses associated with tragic occurrences are commonly the subject of abusive languages. We define a multiclass classification task based on the description above for predicting whether a tweet is normal, hate speech, or abusive language. The dataset used in this study is obtained from the previous research [4] where exhaustive processes of data collection, aggregation, and annotation were conducted. The same dataset was also used in [15], [13], and [7]. Overall, the dataset consists of 13159 tweets written in Indonesian language. We observe that there are 2266 tweets flagged as hate speech, 1748 tweets labelled as abusive language, and 5860 tweets are annotated as normal. The rest of the tweets are labelled as hate speech and abusive language where in this study are omitted. There is a wide range of hate speech classification labels such as hate speech directed at either individual or group, the severity level of hate speech, and the type of hate speech are excluded as the focus of the study is on both hate speech and abusive language [16].

TABLE 1
THE MOST FREQUENT WORDS THAT APPEAR IN EACH LABEL

hate speech	abusive	normal
indonesia (638)	gue (393)	indonesia (704)
jokowi (561)	ya (170)	presiden (691)
tagar (435)	banget (135)	ya (518)
komunis (385)	sih (135)	asing (491)
partai (348)	wkwk (133)	agama (474)
rakyat (277)	kontol (125)	islam (433)
cina (257)	anjir (122)	daerah (412)
presiden (249)	ngentot (106)	gue (400)
islam (246)	kayak (99)	tapi (383)
prabowo (192)	nya (96)	kristen (350)
rezim (177)	memek (95)	gubernur (344)
negara (171)	nih (86)	kepala (322)
lengserkan (167)	tapi (79)	jokowi (317)
korupsi (158)	tai (77)	pilihan (293)
ya (158)	suka (69)	negara (285)
ahok (143)	haha (66)	ekonomi (281)
bubarkan (143)	bodoh (64)	budaya (280)
asing (142)	iya (63)	nya (256)
agama (142)	anjing (59)	katolik (256)
antek (124)	ngewe (58)	komunis (252)

Table 1 shows the 20 most frequent words that appear in each label where highly common words such as pengguna (username), number, tagar (hashtag), and url are removed. It can be seen that frequent words in hate speech are mostly related to politic, race, and religion. In terms of abusive language, most common words are profanity and sex-related terms. Despite containing more general topic, normal tweets are dominated by political terms [16]. In this study we propose a multiclass classification for predicting if a tweet is labelled as normal, hate speech, or abusive language. The tweets go through several preprocessing steps before being fed into the classification models for training. The trained models are then evaluated using accuracy and F1-score metrics. The first step performed in preprocessing is case folding where all characters in tweets are lowercased. This is to allow same words written in uppercase and lowercase to represent the same string. The next process involves cleaning the tweets from characters that are considered as noises such as "RT" (retweet) symbol, emojis, URL, extra spaces, and punctuations. Twitter username, hashtag characters, and numerical characters are replaced with USER, hashtag, and number respectively. The final step is to perform normalization to transform informal words into formal words. This is because many posts on Twitter are written in the informal conversation thus many words are written using informal style. To perform normalization, an informal-to-formal dictionary derived from the previous study in [4] is utilized. Once the data has been preprocessed, the dataset is split into training and testing set. As much as 80% of the dataset is allocated randomly to training set while the rest is used as testing set.

TABLE 2
HATE AND ABUSIVE SPEECH DATASET STATISTICS.

Label	Training Set	Testing Set	Total
Hate Speech	1798 (23%)	468 (24%)	2266
Abusive	1404 (18%)	344 (17%)	1748
Normal	4697 (59%)	1163 (59%)	5860

Table 2 shows the distribution of labels in each training and testing set. Each label are distributed equally between training set and testing set. In terms of modelling, a benchmark model is developed. The benchmark model is

called Majority Class where all tweets in the testing set is labelled as normal. The Majority Class serves as a naïve model that makes no assumption about the problem and its performance is used as a baseline for comparison. In the second model, logistic regression with bag of words technique is developed. The usage of Majority Class and Logistic Regression model with Bag of Words technique were also used in previous similar task in [1] to perform bragging classification. In the logistic regression, one-vs-rest approach is used since this is a multiclass classification problem. In this case, there would be three logistic regression models where each model is trained on one class against the rests. Each model generates a probability score for class membership. The highest-scoring class index is then utilized to predict a class. Equation (1) defines the formula of logistic regression where y is the predicted output being classified into label l among all labels in L and each word x is associated with a weight w [16].

$$P(y = l) = \frac{1}{1 + \exp\left(-(w.x + b)\right)} \quad (1)$$

The other models use Bidirectional Encoder Representations from Transformers (BERT) [11]. Figure 1 shows the overall flow how raw sentences are passed into BERT models and fed into classifier. The BERT model receives tokenized words from sentences and then generate continuous numbers using its pre-trained embeddings. Then, the generated sentence embeddings are fed into the classifier to perform model training. A special [CLS] token is added to the first input token, which stands for Classification. The tokenizer replaces each token with the identifier provided from the embedding from the trained model. BERT receives tokens as input and processes it through a feed-forward network before passing it on to the encoder stack above. Each position outputs a 768 vector of floats. The [CLS] token is used as the input for a classifier in the sentence classification.

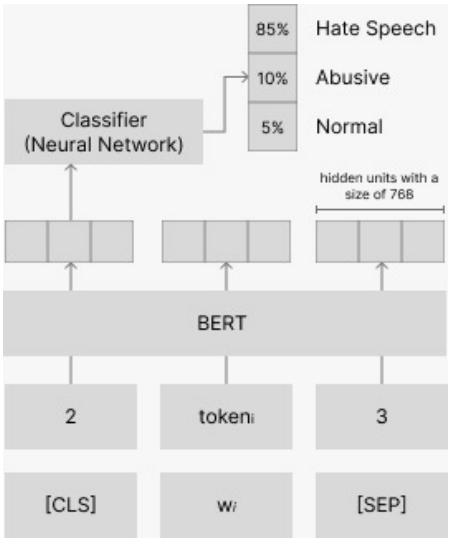

Figure 1. BERT's Input and Output, as Well as Its Classification Flow as Described in [17].

This study experiments with three BERT models of BERT-cahya, BERT-indolem, and BERT-indobechmark which all these three models are Base and uncased model. Generally, there are two versions of BERT models according to its size, the Base and Large version. The total number of encoder layers in Base and Large version is 12 and 24 respectively. The uncased model means that BERT model does not distinguish between lowercase and uppercase wordpiece [16].

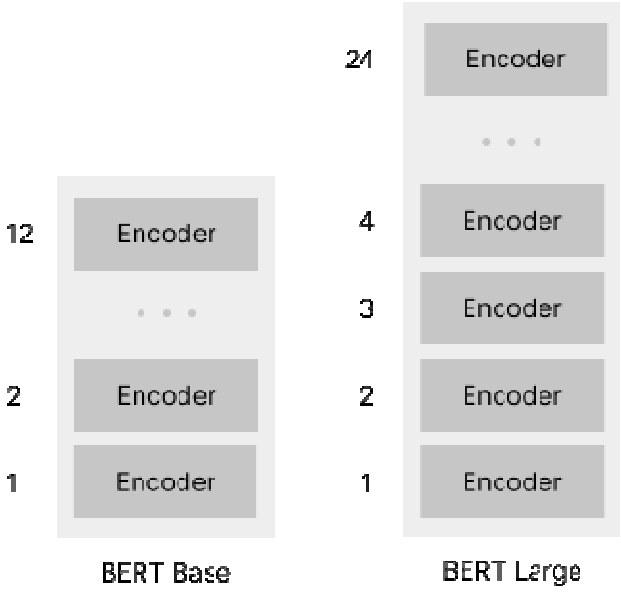

Figure 2. Two Different Versions of BERT by Size, Bert Base and BERT Large [17].

Figure 2 describes two version BERT, BERT Base and BERT Large. BERT Base consists of 12 layers of encoder with 768 hidden unites within its networks and 12 attention heads. In comparison, BERT Large has significantly larger size as there are 24 layers of encoders with 1024 hidden unites and 16 attention heads. The BERT-cahya model is pre-trained with Indonesian Wikipedia while BERT-indolem is pre-trained with Indonesian Wikipedia combined with news articles and Web Corpus. Lastly, BERT-indobenchmark is pre-trained with cleaned dataset gathered from Indonesian public data sources such as social media, news, and blogs [11]. A thorough search of the relevant literature found that these three BERT models have never been used in hate speech detection on Indonesian language [16]. In terms of hyperparameters for these three BERT models, this study uses default parameters from simpletransformers library with the number of training epoch of one and a batch size of eight. Each model is trained and evaluated with several evaluation metrics such as Accuracy, Precision, Recall, and F1-macro.

III. Results and Discussion

Several steps of data processing are carried out to achieve the best possible result.

TABLE 3
DIFFERENT EVALUATION PARAMETERS SCORES FOR EACH MODEL

Model	Accuracy	Precision	Recall	Macro-F1
Majority Class	58.89	19.63	33.33	24.71
LR-BoW	84.15	83.45	81.11	82.19
BERT-cahya	86.53	85.90	**84.22**	84.95
BERT-indolem	83.95	81.87	83.12	82.47
BERT-indobenchmark	**86.99**	**87.69**	83.88	**85.59**

Table 3 demonstrates how well each model predicts if a tweet is normal, hateful, or abusive. In comparison to the Majority Class and Logistic Regression models, BERT models perform better overall. BERT models outperform Majority Class by 60.88 on the Macro-F1-score and Logistic Regression by 3,4 on the Macro-F1-score. In terms of BERT models, BERT-indobenchmark (85.59 F1) achieves better Macro-F1-score than BERT-cahya and BERT-indolem by 0.64 and 3.12 respectively. This shows the significance of pre-training on social media data. As a reminder, both BERT-cahya and BERT-indolem are pre-trained on Indonesian Wikipedia as well as other sources including news articles and web corpus, but not on social media. Although BERT models achieve good overall performance, it is interesting to note that Logistic Regression model achieves a competitive Macro-F1-score performance at 82.19 which is slightly under BERT-indolem at 82.47.

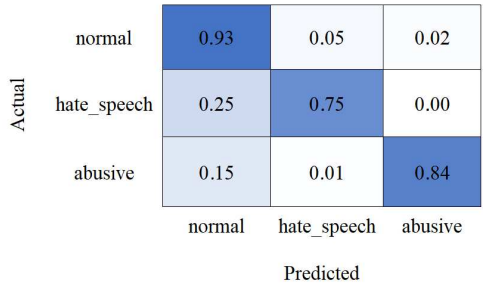

Figure 3. Confusion Matrix of the Best Performing Model – BERT-Indobenchmark.

The confusion matrix of the best performing model BERT-indobenchmark is shown in Figure X. 25% of hate speech tweets and 15% of abusive tweets are misclassified as normal tweets, indicating that the model is likely to categorize other classes as the dominating class. This is most likely due to a class imbalance, as normal tweets account for 59% of the tweets in the dataset. It is also worth noting that the rate of misclassification between hate speech and abusive tweets is quite low, at just above 0%. As a result, it can be observed that the model is capable of distinguishing between hate speech and abusive tweets. It is found that even when normal tweets contain abusive terms or phrases that are frequently linked with negative connotation, they are easily misclassified as hate speech or abusive tweets. For example, Tweet1 and Tweet2 are normal tweets that are classified as hate speech. Both tweets are mentioning words that are often found in hate speech conversation such as criminal and political debate. However, Tweet1 is an invitation to signing a petition regarding an event. Tweet2 is about a somewhat harmless and unbiased viewpoint on communism. Similarly, there are normal tweets that are classified as abusive tweets. As an example, Tweet3 mentions words such as “idiot” and “bodoh” (the Indonesian word for “idiot”) as an idiom which mean the abusive words are used as metaphor with no specific meaning related to insulting others.

Tweet1: tanda tangani petisi ini basuki tjahja purnama tahanan nurani kasus penodaan agama
Tweet2: pengguna pengguna pengguna pengguna anak partai komunis indonesia belum tentu juga partai komunis Indonesia
Tweet3: orang bodoh belum tentu idiot orang idiot belum tentu bodoh haha

Another frequent error happens when hate speech and abusive tweets are categorized as normal. Tweet4 and Tweet5 are hate speech tweets being predicted as normal tweets. The major context of Tweet4 is a threat to stage a protest against a political figure, despite the fact that the amount of threat is rather normal and would not violate the law. Tweet5 is a sarcastic remark towards current political situation in the country. Despite flagged as hate speech, Tweet5 should be labelled as normal tweet and this could be caused by bias during annotation process. In terms of abusive tweet being classified as normal tweet, Tweet6 highlights how words often used in an offensive conversation but used in a harmless context. Tweet6 uses animals with negative characteristic to express the actual animal and not in an offensive manner.

Tweet4: ahok calonkan gubernur siap siap demo
Tweet5: pengguna pengguna kartu kredit tidak punya tetapi pinjaman negara banyak
Tweet6: pengguna monyet bermuka kucing lagi makan pisang

Furthermore, the tweets below shos how some hate speech tweets that was correctly predicted as hate speech by LR-BoW but misclassfied by BERT-indobenchmark and otherwise.

Tweet7: pengguna kalau seiman menipu hukumnya halal tidak mungkin didemo kaum bani micin
Tweet8: bukan kebetulan kalau setelah reformasi bergulir hampir semua media di indonesia dikuasai oleh non muslim moral masyarakat makin hancur

Tweet7 is hate speech but terms used in the sentences are often used in the context of political debate. Therefore LR-BoW is able to predict them as hate speech as it learned from the dataset, making it familiar with those terms. On the other hand, BERT-indobenchmark was trained with more general data gathered from multiple different sources, thus predicting Tweet7 as a normal tweets

In opposite, Tweet 3 is related to hate speech towards certain religion which is a hate speech hate speech and have been predicted correctly by BERT-indobenchmark

Tweet9: besok besok kalau agak budek salahkan pentagon
Tweet10: pak tukang buruan pulang pak saya mau mandi sudah kayak gembel cantik ini

In Tweet9, most of the words contain neutral sentiment, making BERT-indobenchmark to misclassify as a normal tweet while correctly predicted by LR-BoW as an abusive tweet. Tweet10 is predicted correctly by BERT-indobenchmark despite containing positive terms while LR-BoW missclassified it as a normal tweet.

IV. Conclusion

The conclusion of this research is as follows. In conclusion, this research looks on the classification of hate speech, abusive messages, and normal tweets on Indonesian Twitter. A previous study's dataset was utilized, which comprised 5860 tweets categorized as hate speech, abusive, or normal. A benchmark model, logistic regression model, and transformer models of BERT were developed to perform a multiclass classification. Overall, all the machine learning models achieved a high level of accuracy with BERT models slightly outperform logistic regression. The best performing model of BERT-indobenchmark has the Macro-F1-score at 85.59. This demonstrates that the model can distinguish between hate speech and abusive language on Indonesian Twitter. It's critical to distinguish between hate speech and abusive tweets as hate speech frequently results in legal action. Separating hate speech from abusive language has the benefit of allowing people to express themselves without fear of facing legal consequences. People on social media, particularly Indonesian Twitter users, would benefit the most in this scenario since they are able to distinguish between hate speech and non-hate speech posts. As a result, they are protected from legal action.

In the future, the dataset for such research should include a broader range of tweets. This is because numerous terms present in abusive, hate speech, and normal tweets have been found to be dominated by tweets related to political events. As a result, collecting more general data could reduce any potential bias. Another potential work in the future is to use the Explainable AI approach to complement the hate speech classification model. This approach would provide explanation as to how certain predictions are classified into specific labels.

Acknowledgment

The researcher wishes to express his gratitude to the experts for their helpful and constructive ideas for improving this study. Bina Nusantara University was also thanked. This research was made possible thanks to the Binus Initiative Project grant's facilities, finances, and infrastructure.

References

[1] G. H. Stanton and G. H. Stanton, "Journal of African Conflicts and Peace Studies The Rwandan Genocide : Why Early Warning Failed," vol. 1, no. 2, pp. 6–25, 2009.

[2] A. A. Abdillah, Azwardi, S. Permana, I. Susanto, F. Zainuri, and S. Arifin, "Performance Evaluation Of Linear Discriminant Analysis And Support Vector Machines To Classify Cesarean Section," *Eastern-European J. Enterp. Technol.*, vol. 5, no. 2–113, pp. 37–43, 2021, doi: 10.15587/1729-4061.2021.242798.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017*, no. Icwsm, pp. 512–515, 2017.

[4] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," pp. 46–57, 2019, doi: 10.18653/v1/w19-3506.

[5] S. Arifin *et al.*, "Graph Coloring Program of Exam Scheduling Modeling Based on Bitwise Coloring Algorithm Using Python," 2022, doi: 10.3844/jcssp.2022.26.32.

[6] H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal, "Deep learning for detecting inappropriate content in text," *Int. J. Data Sci. Anal.*, vol. 6, no. 4, pp. 273–286, 2018, doi: 10.1007/s41060-017-0088-4.

[7] R. Hendrawan, "Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media".

[8] 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)*, 2017, pp. 233–238. doi: 10.1109/ICACSIS.2017.8355039.

[9] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," *CEUR Workshop Proc.*, vol. 1816, pp. 86–95, 2017.

[10] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," *25th Int. World Wide Web Conf. WWW 2016*, pp. 145–153, 2016, doi: 10.1145/2872427.2883062.

[11] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media," *Stud. Comput. Intell.*, vol. 881 SCI, pp. 928–940, 2020, doi: 10.1007/978-3-030-36687-2_77.

[12] J. A. Leite, D. F. Silva, K. Bontcheva, and C. Scarton, "Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis," 2020.

[13] S. Kurniawan and I. Budi, "Indonesian Tweets Hate Speech Target Classification using Machine Learning," pp. 1–5.

[14] K. J. Madukwe, X. Gao, and B. Xue, "Dependency-based embedding for distinguishing between hate speech and offensive language," *Proc. - 2020 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. WI-IAT 2020*, pp. 860–868, 2020, doi: 10.1109/WIIAT50758.2020.00132.

[15] M. O. Ibrohim, "Identification of Hate Speech and Abusive Language on Indonesian Twitter Using theWord2vec , Part of Speech and Emoji Features," 2019.

[16] M. A. Ibrahim *et al.*, "AN EXPLAINABLE AI MODEL TO HATE SPEECH DETECTION ON INDONESIAN TWITTER," *CommIT (Communication Inf. Technol. J.*, vol. 16, no. 2, 2022.

[17] D. Fimoza, A. Amalia, and T. H. F. Harumy, "Sentiment Analysis for Movie Review in Bahasa Indonesia Using BERT," pp. 27–34, 2021.