International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 99, No. 1, Month 2099, pp. 1~1x
ISSN: 2252-8776, DOI: 10.11591/ijict.v99i1.pp1-1x

MULTILINGUAL HATE SPEECH DETECTION USING DEEP LEARNING

Vincent, Amalia Zahra
Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Bi-LSTM
Data Augmentation
Hate Speech
M-BERT
Natural Language Processing

ABSTRACT

The rise of social media has enabled public expression but also fueled the spread of hate speech, contributing to social tensions and potential violence. Natural Language Processing (NLP), particularly text classification, has become essential for detecting hate speech. This study develops a hate speech detection model on Twitter using FastText with Bidirectional Long Short-Term Memory (Bi-LSTM) and explores multilingual Bidirectional Encoder Representations from Transformers (M-BERT) for handling diverse languages. Data augmentation techniques—including Easy Data Augmentation (EDA) methods, Back Translation, and Generative Adversarial Networks (GANs)—are employed to enhance classification, especially for imbalanced datasets. Results show that data augmentation significantly boosts performance. The highest F1-scores are achieved by Random Insertion for Indonesian (F1-score: 0.889, Accuracy: 0.879), Synonym Replacement for English (F1-score: 0.872, Accuracy: 0.831), and Random Deletion for German (F1-score: 0.853, Accuracy: 0.830) with the FastText + Bi-LSTM model. The M-BERT model performs best with Random Deletion for Indonesian (F1-score: 0.898, Accuracy: 0.880), Random Swap for English (F1 score: 0.870, Accuracy: 0.866), and Random Deletion for German (F1-score: 0.662, Accuracy: 0.858). These findings underscore that data augmentation effectiveness varies by language and model. This research supports efforts to mitigate hate speech's impact on social media by advancing multilingual detection capabilities.

Corresponding Author:

Vincent
Computer Science Department, Binus Graduate Program
Master of Computer Science, Bina Nusantara University,
Jakarta, Indonesia
Email: vincent053@binus.ac.id

1. INTRODUCTION

Social media is a part of the internet that facilitates users in expressing themselves, collaborating, interacting, sharing, and communicating with others. In 2020, half of the world's population was already using social media. Out of 5.2 billion people who owned mobile phones, 4.5 billion were connected to the internet, and 3.8 billion were active social media users. This figure of 3.8 billion represented 49% of the global population in 2020. The emergence of social media platforms like Twitter, Facebook, and Instagram has provided a space for the public to express their opinions through posted comments. Among these platforms, Twitter stands out by offering features like retweets, likes, and replies, allowing millions of users to indirectly interact in conversations. However, these features have introduced a significant new challenge: the spread of hate speech [1].

Hate speech is a phenomenon where users, either intentionally or unintentionally, disseminate messages or content aimed at provoking, inciting, or discriminating against specific individuals or groups [2]. Hate speech often targets individuals based on characteristics such as ethnicity, religion, race, gender, sexual orientation, or other backgrounds. Consequently, hate speech can lead to social tensions, create stereotypes, harm intercultural relations, and incite other forms of physical violence. To address this issue, research in Natural Language Processing (NLP), particularly in text classification, has emerged as an effective approach to analyzing the complex linguistic patterns used on social media. Text classification is a branch of NLP aimed at grouping texts into certain categories or classes. However, challenges in text classification arise when the model must handle the linguistic and contextual diversity on social media, including unique writing styles and phrases [3]. Additionally, noisy and imbalanced data on social media present further challenges for text classification [4], [5]. This noise can stem from various factors, including abbreviations, slang, emoticons, or other writing errors. Meanwhile, imbalanced data can cause issues, such as overfitting to the majority class and poor predictive accuracy.

To address these challenges, the development of models capable of handling linguistic variation and classification challenges is the focus of this research. Methods such as data augmentation, adaptive natural language processing, and deep learning approaches have become increasingly popular solutions for tackling these challenges. With technological advancements and more sophisticated methods, text classification is expected to continually improve in its performance and accuracy in interpreting the diverse information found on social media. In this study, the authors will develop a model using fastText combined with Bidirectional Long Short-Term Memory (Bi-LSTM) to analyze public opinion on Twitter regarding various issues. The selection of the Bi-LSTM model is based on its ability to handle computational challenges and enhance a holistic understanding of text classification [6]. This research will also explore the potential for transfer learning development using multilingual Bidirectional Encoder Representations from Transformers (M-BERT) to improve model performance in detecting and addressing multilingual hate speech. Additionally, data augmentation methods such as Easy Data Augmentation (EDA), Back Translation, and Generative Adversarial Networks (GANs) will be utilized to enhance model performance.

This study is expected to contribute to efforts in minimizing the negative impact of hate speech in social media. The research paper will be structured as follows. Section 2 will present related previous work on multilingual methods from various sources. Section 3 will discuss the proposed method used in this research. The results of the model development will be displayed through tables in Section 4, with a concluding summary provided in Section 5.

2. LITERATURE REVIEW

Research utilizing multilingual learning models, such as Multilingual BERT (mBERT), has been conducted for some time. In the study [7], Hate Speech classification achieved notable results using mBERT in various settings. For the multilingual baseline mBERT, the model showed moderate accuracy, with the highest performance observed in Hindi (73.7%), followed by English (73.6%), and lower accuracy in German (69.9%). These results indicate that the baseline mBERT model can handle multilingual Hate Speech classification reasonably well but struggles with certain languages like German. When augmented with Generative Adversarial Networks (GAN), significant improvements were observed across all languages. The accuracy for Hindi increased to 78.3%, while English reached 75.3%, and German showed the most substantial improvement, rising to 77.1%. These findings highlight the effectiveness of combining mBERT with GAN to enhance Hate Speech classification, particularly in addressing language diversity and boosting classification accuracy across multilingual datasets.

Additionally, researchers [8] compared different word embeddings, such as MUSE and ELMO, to assess their influence on model performance. However, the M-BERT model does not always outperform other models in terms of F1-score. For instance, in sentiment analysis research by [9], fastText + Bi-LSTM ranked second, closely competing with M-BERT, with F1-scores of 80.37% on the UCSA dataset and 76.50% on the UCSA-21 dataset. In this study, the use of data augmentation with Generative Adversarial Networks (GANs) proved to enhance accuracy by 6%, showing a superior performance compared to back-translation and easy data augmentation [10].

3. METHOD

In this section, we describe the research methodology, focusing on text classification techniques outlined in Figure 1. The process began with data collection, followed by preprocessing to clean, normalize, and structure the data, ensuring consistency and quality across samples. The dataset was then divided into training, validation, and testing sets to enable a comprehensive evaluation of model performance. Various data augmentation techniques were subsequently applied to expand and diversify the dataset, enhancing model generalizability. Next, the model was trained using optimized hyperparameters selected through systematic tuning for improved performance. Finally, each model's effectiveness and accuracy were rigorously evaluated and analyzed within the context of text classification for the chosen domain.

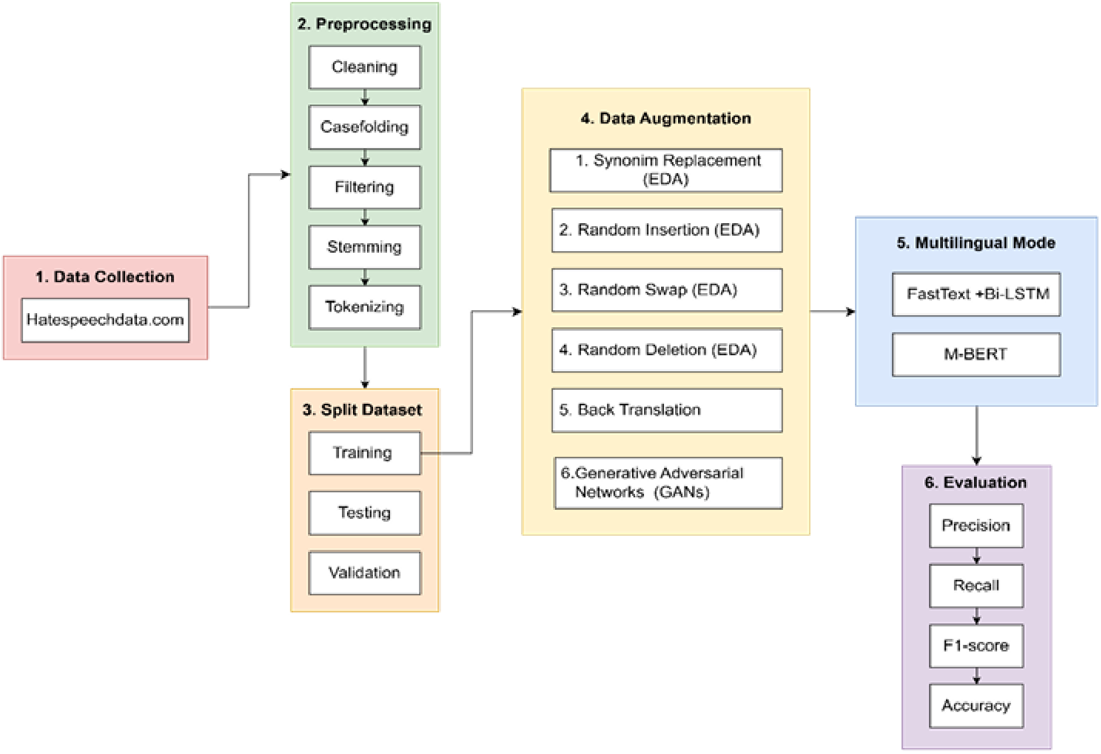


Figure 1. Proposed Method

3.1. Data Collection

The data for this research was collected from Hatespeechdata.com, a reliable platform providing datasets for hate speech analysis. The dataset used in this study includes text in three languages: Indonesian, English, and German. Each piece of text in the dataset is labeled to indicate whether it constitutes hate speech or non-hate speech. Non-hate speech instances are labeled with a binary 0, while hate speech instances are labeled with a binary 1. As an example, this can be seen in Table 1.

Table 1. Example Data Collection

Tweet	Language	Label
Hehehe itulah dungunya kamu otak terbalik susah untuk dijelaskan itu anak ingusan baru tumbuh belum pantas mengerti sana nete dulu sama ibumu Dasar kadal kebon	Indonesia	1
terima kasih pak kini papua punya wajah baru gak meninggalkan budaya mereka tapi pembangunan tidak mengabaikan kearifan local	Indonesia	0
avoid political posts but FUCK You Mr. Modi, fuck your politics and all your inhumane tactics to let people die. https://t.co/hGagTj3IIu #IndiaCovidCrisis	English	1
@martin_compston @BritBox_US Ã¢â,¬Å“Mary , Joseph and the wee Donkey, have you not heard of photoshop son? You've just at the internet a challenge, a challenge they'll take!"	English	0
RT @Hihi97948034: Mero ist ein Hurensohn	German	1
Barkley > Xavi https://t.co/SAncB0kT6F	German	0

3.2. Preprocessing Data

Text preprocessing involves a series of steps to manipulate and clean text for more effective processing with natural language processing. Preprocessing includes several stages to clean and transform raw text into a format suitable for analysis in modeling, ensuring high-quality results The following are the preprocessing steps [11]:

1. Cleaning: The process of removing special characters, numbers, spaces, URLs, emoticons, and unnecessary symbols from the text. The main goal is to clean the text of irrelevant elements or noise.
2. Case Folding: Converting all letters in the text to either lowercase or uppercase.
3. Filtering: Irrelevant words, particularly stop words, are removed from the text. Stop words are commonly used terms that provide little value to the analysis, such as "the," "is," or "and" and filtering them out helps focus on more relevant words.
4. Stemming: The process of removing word suffixes to return the word to its root form. Specifically, the method optimizes analysis by handling different word forms.
5. Tokenizing: Breaking the text into smaller units such as words or phrases.

3.3. Split Dataset

In data processing for deep learning, it is important to split the data into subsets so that the model can be effectively training, validation, and testing.

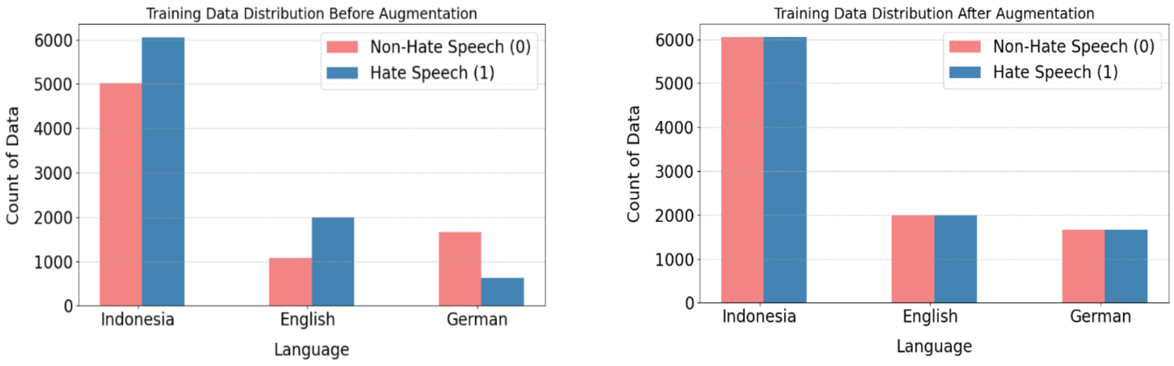
Table 2. Splitting Data

Language	Label	Training (80%)	Validation (10%)	Testing (10%)
Indonesia	Non-Hate Speech (0)	5,023	628	628
	Hate Speech (1)	6,048	756	756
English	Non-Hate Speech (0)	1,073	134	134
	Hate Speech (1)	2,001	250	250
German	Non-Hate Speech (0)	1,659	207	208
	Hate Speech (1)	633	79	79

From Table 2, we see that a common approach is to divide the data into 80% training data, 10% validation data, and 10% testing data. This split ensures that the model has enough data to learn from (training data), monitor and adjust its performance (validation data), and measure its final performance on previously unseen data (testing data). With this allocation, the model is expected to generalize well and deliver accurate performance.

3.4. Data Augmentation

Data augmentation techniques are essential for addressing the imbalance in the dataset, as shown in Table 1 and Figure 2. In the Indonesian dataset, the number of non-hate speech samples (5,023) is significantly lower than that of hate speech samples (6,048), with a difference of 1,025 instances. A similar imbalance is seen in the English dataset, where non-hate speech (1,073) is considerably less than hate speech (2,001), resulting in a difference of 928 instances. The German dataset shows the most extreme imbalance, with non-hate speech (1,659) outnumbering hate speech (633) by 1,026 instances.This imbalance can limit the model’s ability to generalize and introduce bias toward the majority class, which in turn can reduce the accuracy of predictions for the minority class.



(a) (b)

Figure 2. Training Data (a) Before Augmentation and (b) After Augmentation

Figure 2 illustrates the distribution of training data across three languages: Indonesian, English, and German, comparing the data before and after augmentation. Subfigure 2(a) represents the count of data samples in the non-augmented dataset. Subfigure 2(b) shows the distribution after applying data augmentation techniques, which resulted in a more balanced representation of hate speech (labeled as "1") and non-hate speech (labeled as "0") across the three languages. To achieve this balance, the study employs three data augmentation techniques: Easy Data Augmentation (EDA), Back Translation, and Generative Adversarial Networks (GANs). These methods enhance the diversity and volume of training data, addressing the imbalance and improving the model's ability to generalize across multilingual datasets.

3.4.1 Easy Data Augmentation

EDA is a technique used in Natural Language Processing (NLP) to increase the amount of training data by generating variations of existing data . The primary goal of EDA is to improve model performance by introducing slight variations in the training data, enabling the model to learn from a wider range of examples. The main types of Easy Data Augmentation (EDA) include [12], [13]:

1. Synonym Replacement: In this approach, words in the text are replaced with their synonyms. This can be done using a thesaurus or word embedding models to find words with similar meanings in specific contexts.
2. Random Insertion: This method involves randomly inserting additional words into the text. These extra words can be chosen from other data This helps introduce more variety and richness in the data.
3. Random Swap: In this approach, the order of words in a sentence is randomly swapped to create new variations of the text.
4. Random Deletion: This method involves randomly deleting words from the text. This introduces variation by removing certain words, forcing the model to rely more on the context of the remaining words.

3.4.2 Back Translation

Back Translation is used as a data augmentation technique in Natural Language Processing to increase variation and the amount of training data. This technique involves translating the original text into another language and then translating it back to the original language [14]. By doing so, it introduces subtle changes in sentence structure, vocabulary, and phrasing, which help the model generalize better to unseen data. This approach is particularly effective for low-resource languages or datasets with limited diversity. In this study, Indonesian, English, and German data will be translated into French and then back to their respective original languages. The use of French as an intermediate language was chosen to leverage its structural differences from the source languages, thereby introducing meaningful variations. This process not only enriches the training dataset but also enhances the robustness of models by exposing them to multiple linguistic patterns and variations derived from translation.

3.4.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are an innovative method in text data augmentation, consisting of two main components: the generator and the discriminator [15]. The generator is responsible for creating synthetic text that resembles the original text, aiming to mimic the linguistic patterns and semantic structure found in the training data. On the other hand, the discriminator acts as a supervisor, evaluating and distinguishing between the original text and the synthetic text generated by the generator. During the training process, the generator continuously improves its ability to produce more realistic and contextually appropriate text, while the discriminator becomes increasingly skilled at identifying subtle differences between real and synthetic text. This adversarial training dynamic pushes both components to improve iteratively, ultimately resulting in a generator that can produce highly convincing synthetic text. By leveraging this method, GANs can effectively enhance datasets, improve model robustness, and address data scarcity challenges in tasks such as sentiment analysis, machine translation, and text classification [16] .

3.5. Multilingual Model

3.5.1 FastText +Bi-LSTM

FastText is a word representation model developed by Facebook AI Research (FAIR) that is used as word embedding for the three languages in this study: Indonesian, English, and German [17], [18]. To maximize the use of FastText word representations, this model is combined with Bi-LSTM (Bidirectional Long Short-Term Memory), a type of RNN designed to handle long-range dependencies in sequential data. Bi-LSTM processes the input not only in the forward direction but also in the backward direction, allowing the model to capture context from both sides of a token in the sequence [19], [20]. The combination of FastText for word representation and Bi-LSTM for sequence modeling enables the model to better understand the semantic meaning of text and its context, making it suitable for text classification tasks such as hate speech detection . In this study, extensive hyperparameter tuning was conducted to identify the optimal settings for the combined FastText and Bi-LSTM model. Key parameters adjusted included the learning rate, dropout rate, and the number of units in the Bi-LSTM layer. Each configuration was evaluated based on validation accuracy to determine its effectiveness in capturing contextual and semantic information for accurate text classification. This process is summarized in Table 2.

Table 3. Best Hyperparameter Tuning

Hyperparameter			Validation Accuracy
Learning Rate	Dropout Rate	Units	
0.01	0.5	32	0.8364
0.01	0.5	8	0.8359
0.01	0.4	32	0.8354
0.01	0.4	16	0.8349
0.001	0.4	8	0.8349
0.01	0.4	16	0.8349

After conducting various hyperparameter experiments and analyzing the training and validation accuracy curves, the optimal configuration in table 3 was selected with a learning rate of 0.001, dropout of 0.4, and 8 units. This configuration was chosen because the accuracy curves demonstrate minimal overfitting, with a stable relationship between training and validation accuracy.

3.5.2 M-BERT

M-BERT (Multilingual BERT) is a transformer-based model designed to process text in multiple languages simultaneously [21], [22]. It is the multilingual version of BERT, trained on text data from various languages without separating models by language. With pre-training and fine-tuning, M-BERT captures semantic meaning and syntactic structure from text, leveraging both previous and following words through its transformer architecture. The M-BERT model used in this research is from the google-bert/bert-base-multilingual-cased version, enabling it to process data in Indonesian, English, and German. This model [23], [24] is particularly beneficial for tasks involving multilingual datasets, as it ensures consistent and robust performance across languages without requiring language-specific adaptations. By using shared parameters across languages, M-BERT effectively learns multilingual representations

4. RESULT AND DISCUSION

In this section, we present the results and discussion of our research, focusing on the evaluation of model performance using accuracy and F1-score metrics.

Table 4. Model Comparison FastText + Bi-LSTM and M-BERT

Model	Data Augmentation	Indonesia		English		German	
		Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
FastText+ Bi-LSTM	None (Baseline)	0.859	0.873	0.779	0.838	0.804	0.575
	EDA (Synonym Replacement)	0.875	0.888	**0.831**	**0.872**	0.808	0.670
	EDA (Random Insertion)	**0.879**	**0.889**	0.813	0.850	0.853	0.720
	EDA (Random Swap)	0.865	0.873	0.794	0.831	**0.881**	0.763
	EDA (Random Deletion)	0.869	0.877	0.792	0.830	0.853	**0.853**
	Back Translation	0.872	0.885	0.815	0.863	0.832	0.684
	GANs	0.826	0.862	0.758	0.824	0.818	0.623
M-BERT	None (Baseline)	**0.886**	0.897	0.776	0.829	0.787	0.590
	EDA (Synonym Replacement)	0.876	0.889	0.815	0.864	**0.822**	0.653
	EDA (Random Insertion)	0.864	0.879	**0.818**	0.861	0.818	0.638
	EDA (Random Swap)	0.865	0.883	0.815	**0.866**	0.780	**0.663**
	EDA (Random Deletion)	0.880	**0.898**	0.802	0.858	0.818	0.662
	Back Translation	0.872	0.887	0.801	0.855	0.780	0.577
	GANs	0.867	0.883	0.722	0.807	0.770	0.609

Table 5. Average Model Comparison

Model	Data Augmentation	Indonesia		English		German	
		Accuracy	F1-Score	Accuracy	F1-score	Accuracy	F1-score
FastText+ Bi-LSTM	None (Baseline)	0.859	0.873	0.779	0.838	0.804	0.575
	Average with 6 data augmentation	0.864	0.879	**0.800**	0.845	**0.840**	**0.718**
M-BERT	None (Baseline)	**0.886**	**0.897**	0.776	0.829	0.787	0.590
	Average with 6 data augmentation	0.870	0.886	0.795	**0.851**	0.798	0.633

In the Model Evaluation, we thoroughly assess the performance of the proposed model using two key evaluation metrics, as referenced in prior studies [25], [26]:

1. Accuracy is a metric that calculates the proportion of correctly classified instances out of the total instances in the dataset. It is determined by dividing the sum of true positive and true negative predictions by the total number of instances, as shown in the formula:

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

2. F1-Score is the harmonic mean of Precision and Recall, and it provides a balanced measure of a model's accuracy by combining both metrics. The F1-score is particularly useful when dealing with imbalanced datasets, as it considers both the precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positives among all actual positives). The formula for F1-score is:

$$F1 - Score = \frac{2 \times precision \times Recall}{Precision + Recall} \tag{2}$$

A higher accuracy value indicates that the model is effective in correctly identifying both positive and negative classes, providing a straightforward measure of its overall performance. Besides that, a higher F1-score indicates better performance, especially in scenarios where certain classes may dominate. This makes F1-score a valuable metric when accuracy alone may not provide a complete model’s effectiveness.

Based on the results in Table 4, we identify the best-performing data augmentation techniques for each language across both models. For the FastText + Bi-LSTM model, *Random Insertion* performs best for Indonesian with an accuracy of 0.879 and an F1-score of 0.889, *EDA (Synonym Replacement)* is most effective for English with an accuracy of 0.831 and an F1-score of 0.872, and *Random Deletion* yields the highest performance for German with an accuracy of 0.830 and an F1-score of 0.853. In the case of M-BERT, *Random Deletion* shows the best results for Indonesian (accuracy: 0.880, F1-score: 0.898), *Random Swap* is optimal for English (accuracy: 0.866, F1-score: 0.870), and *Random Deletion* again achieves the highest accuracy and F1-score for German (accuracy: 0.858, F1-score: 0.662). These findings suggest that the effectiveness of data augmentation techniques varies by language and model.

Table 5 provides a comparison of average model performance with and without data augmentation for both FastText + Bi-LSTM and M-BERT across three languages. For FastText + Bi-LSTM, the inclusion of six data augmentation techniques improves performance for English and German, yielding higher accuracy (0.800 for English and 0.840 for German) and F1-scores (0.845 for English and 0.718 for German) compared to the baseline. However, for Indonesian, the accuracy (0.864) and F1-score (0.879) are only slightly higher than the baseline.

In the case of M-BERT, applying data augmentation also enhances performance, especially for English, where the F1-score improves from 0.829 to 0.851. The accuracy and F1-scores for German also increase with data augmentation, though the improvement is more modest (from 0.787 to 0.798 in accuracy and from 0.590 to 0.633 in F1-score). For Indonesian, however, the baseline model (accuracy of 0.886 and F1-score of 0.897) slightly outperforms the average with data augmentation (accuracy of 0.870 and F1-score of 0.886). Overall, data augmentation generally improves performance, with the most notable gains observed for English and German, especially in terms of F1-score, while the impact on Indonesian is less significant.

5. CONCLUSION

In conclusion, our research demonstrates the impact of various data augmentation techniques on multilingual hate speech detection across Indonesian, English, and German datasets. By leveraging augmentation methods like random insertion, synonym replacement, and random deletion, we observed notable improvements in model performance, particularly with F1-scores, which are crucial in imbalanced datasets typical of hate speech detection. The FastText + Bi-LSTM and M-BERT models showed differing strengths depending on the language, highlighting that language-specific nuances can significantly influence the effectiveness of both models and augmentation strategies. English and German datasets benefited the most from augmentation, with substantial boosts in both accuracy and F1-scores, whereas the Indonesian dataset exhibited only marginal improvements, suggesting it may capture language nuances effectively even without augmentation.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Vincent	✓	✓	✓	✓		✓		✓	✓	✓	✓		✓	
Amalia Zahra	✓	✓				✓		✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

REFERENCES:

[1] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López, "Internet, social media and online hate speech. Systematic review," *Aggress Violent Behav*, vol. 58, p. 101608, May 2021, doi: 10.1016/j.avb.2021.101608.

[2] M. A. Paz, J. Montero-Díaz, and A. Moreno-Delgado, "Hate Speech: A Systematized Review," *Sage Open*, vol. 10, no. 4, Oct. 2020, doi: 10.1177/2158244020973022.

[3] P. Jotikabukkana, V. Sornlertlamvanich, O. Manabu, and C. Haruechaiyasak, "Social Media Text Classification by Enhancing Well-Formed Text Trained Model," *Journal of ICT Research and Applications*, vol. 10, no. 2, pp. 177–196, Aug. 2016, doi: 10.5614/itbj.ict.res.appl.2016.10.2.6.

[4] B. Liu and Y. Wang, "Deep Learning Models for Text Classification," in *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, IEEE, Dec. 2022, pp. 821–826. doi: 10.1109/TOCS56154.2022.10015969.

[5] PM. Lavanya and E. Sasikala, "Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, IEEE, May 2021, pp. 603–609. doi: 10.1109/ICSPC51351.2021.9451752.

[6] B. Jang, M. Kim, G. Harerimana, S. Kang, and J. W. Kim, "Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism," *Applied Sciences*, vol. 10, no. 17, p. 5841, Aug. 2020, doi: 10.3390/app10175841.

[7] K. Mnassri, R. Farahbakhsh, and N. Crespi, "Multilingual Hate Speech Detection: A Semi-Supervised Generative Adversarial Approach," *Entropy*, vol. 26, no. 4, p. 344, Apr. 2024, doi: 10.3390/e26040344.

[8] M. Bojkovský and M. Pikuliak, "STUFIIT at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter with MUSE and ELMo Embeddings," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 464–468. doi: 10.18653/v1/S19-2082.

[9] L. Khan, A. Amjad, N. Ashraf, and H.-T. Chang, "Multi-class sentiment analysis of urdu text using multilingual BERT," *Sci Rep*, vol. 12, no. 1, p. 5436, Mar. 2022, doi: 10.1038/s41598-022-09381-9.

[10] P. Ghadekar, M. Jamble, A. Jaybhay, B. Jagtap, A. Joshi, and H. More, "Text Data Augmentation Using Generative Adversarial Networks, Back Translation and EDA," 2023, pp. 391–401. doi: 10.1007/978-3-031-37940-6_32.

[11] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 1, p. 776, Feb. 2022, doi: 10.11591/ijece.v12i1.pp776-784.

[12] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Appl Soft Comput*, vol. 132, p. 109803, Jan. 2023, doi: 10.1016/j.asoc.2022.109803.

[13] A. M. Issifu and M. C. Ganiz, "A Simple Data Augmentation Method to Improve the Performance of Named Entity Recognition Models in Medical Domain," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, IEEE, Sep. 2021, pp. 763–768. doi: 10.1109/UBMK52708.2021.9558986.

[14] D. R. Beddiar, M. S. Jahan, and M. Oussalah, "Data expansion using back translation and paraphrasing for hate speech detection," *Online Soc Netw Media*, vol. 24, p. 100153, Jul. 2021, doi: 10.1016/j.osnem.2021.100153.

[15] G. Iglesias, E. Talavera, and A. Díaz-Álvarez, "A survey on GANs for computer vision: Recent research, analysis and taxonomy," *Comput Sci Rev*, vol. 48, p. 100553, May 2023, doi: 10.1016/j.cosrev.2023.100553.

[16] L. Gonog and Y. Zhou, "A Review: Generative Adversarial Networks," in *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, Jun. 2019, pp. 505–510. doi: 10.1109/ICIEA.2019.8833686.

[17] M. Umer *et al.*, "Impact of convolutional neural network and FastText embedding on text classification," *Multimed Tools Appl*, vol. 82, no. 4, pp. 5569–5585, Feb. 2023, doi: 10.1007/s11042-022-13459-x.

[18] A. Amalia, O. S. Sitompul, E. B. Nababan, and T. Mantoro, "An Efficient Text Classification Using fastText for Bahasa Indonesia Documents Classification," in *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, IEEE, Jul. 2020, pp. 69–75. doi: 10.1109/DATABIA50434.2020.9190447.

[19] A. Glenn, P. LaCasse, and B. Cox, “Emotion classification of Indonesian Tweets using Bidirectional LSTM,” *Neural Comput Appl*, vol. 35, no. 13, pp. 9567–9578, May 2023, doi: 10.1007/s00521-022-08186-1.

[20] U. B. Mahadevaswamy and P. Swathi, “Sentiment Analysis using Bidirectional LSTM Network,” *Procedia Comput Sci*, vol. 218, pp. 45–56, 2023, doi: 10.1016/j.procs.2022.12.400.

[21] T. Pires, E. Schlinger, and D. Garrette, “How Multilingual is Multilingual BERT?,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4996–5001. doi: 10.18653/v1/P19-1493.

[22] P. Wu, X. Li, C. Ling, S. Ding, and S. Shen, “Sentiment classification using attention mechanism and bidirectional long short-term memory network,” *Appl Soft Comput*, vol. 112, p. 107792, Nov. 2021, doi: 10.1016/j.asoc.2021.107792.

[23] A. N. Azhar and M. L. Khodra, “Fine-tuning Pretrained Multilingual BERT Model for Indonesian Aspect-based Sentiment Analysis,” in *2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)*, IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ICAICTA49861.2020.9428882.

[24] G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis, “Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data,” *Neural Comput Appl*, vol. 35, no. 29, pp. 21415–21431, Oct. 2023, doi: 10.1007/s00521-023-08629-3.

[25] Y. Liu and S. Yang, “Application of Decision Tree-Based Classification Algorithm on Content Marketing,” *Journal of Mathematics*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/6469054.

[26] Ž. Đ. Vujovic, “Classification Model Evaluation Metrics,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.

BIOGRAPHIES OF AUTHORS

Vincent is a student pursuing a bachelor’s degree in computer science at Bina Nusantara University. He has demonstrated passion for Natural Language Processing. This publication marks his inaugural contribution to the academic community and reflects eagerness to engage with scholarly discourse. He can be contacted at email: vincent053@binus.ac.id.

Amalia Zahra, S.Kom., Ph. D. has been a lecturer at the Master of Computer Science department, Bina Nusantara University since 2017. She earned her Ph.D. from the School of Computer Science and Informatics at University College Dublin (UCD), Ireland, in 2014. She completed her bachelor’s degree at the Faculty of Computer Science, University of Indonesia (UI), in 2008. She gained significant research experience as a research assistant working on the development of an Indonesian Speech Recognition System after graduating from UI. Before joining Bina Nusantara University, she was a lecturer and researcher at the Faculty of Computer Science, University of Indonesia. Her research interests encompass Speech Processing, Speech Recognition, Speaker Recognition, Spoken Language Identification, and other areas related to Speech Technology. She can be contacted at email: amalia.zahra@binus.edu.