Journal of Computer Science

Original Research Paper

Exploring Data Augmentation for Gender-Based Hate Speech Detection

[1]Muhammad Amien Ibrahim, [2]Samsul Arifin and [1]Eko Setyo Purwanto

[1]Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[2]Department of Statistics, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article history
Received: 14-04-2023
Revised: 26-07-2023
Accepted: 03-08-2023

Corresponding Author:
Muhammad Amien Ibrahim
Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
Email: muhammad.amien@binus.ac.id

Abstract: Social media moderation is a crucial component to establish healthy online communities and ensuring online safety from hate speech and offensive language. In many cases, hate speech may be targeted at specific gender which could be expressed in many different languages on social media platforms such as Indonesian Twitter. However, difficulties such as data scarcity and the imbalanced gender-based hate speech dataset in Indonesian tweets have slowed the development and implementation of automatic social media moderation. Obtaining more data to increase the number of samples may be costly in terms of resources required to gather and annotate the data. This study looks at the usage of data augmentation methods to increase the amount of textual dataset while keeping the quality of the augmented data. Three augmentation strategies are explored in this study: Random insertion, back translation, and a sequential combination of back translation and random insertion. Additionally, the study examines the preservation of the increased data labels. The performance result demonstrates that classification models trained with augmented data generated from random insertion strategy outperform the other approaches. In terms of label preservation, the three augmentation approaches have been shown to offer enough label preservation without compromising the meaning of the augmented data. The findings imply that by increasing the amount of the dataset while preserving the original label, data augmentation could be utilized to solve issues such as data scarcity and dataset imbalance.

Keywords: Dataset, Data Augmentation, Hate Speech Detection

Introduction

The use of social media as a communication tool has become critical in today's culture. Nonetheless, it is common to encounter hate speech on these platforms, which has an impact on everyone's safety as well as society as a whole. The increase in the use of abusive and offensive language on social media platforms such as Twitter has had an impact on a wide spectrum of conversations, including political, racial, religious, and gender discourse in a variety of languages. As a result, social media moderation is essential to detect and prevent hate speech from spreading. However, an automated tool to detect and execute moderation is required due to the overwhelming volume of social media postings to be monitored. Previous research on automated hate speech detection has covered a wide range of languages, including English (Meng *et al*., 2023; Mozafari *et al*., 2020; Lin *et al*., 2021), Spanish (Plaza-Del-Arco *et al*., 2020; 2021; García-Díaz *et al*., 2023; Pereira-Kohatsu *et al*., 2019), Indonesian (Ibrahim *et al*, 2022a; Hendrawan and Al Faraby, 2020; Ibrohim *et al*., 2019) and Urdu (Amjad *et al*., 2021). Hate speech detection is a tool to determine whether a specific post is generally hate speech or not. Another study on hate speech detection attempts to establish at a granular level, for example, whether a hate speech post is intended toward women (Basile *et al*., 2019). This would allow hate speech detection to be more exact in identifying hate speech. However, developing a more granular hate speech detection tool requires a dataset that allows for data training with its matching label at the granular hate speech level. For instance, a gender-based hate speech dataset would require text samples with matching gender-based hate speech labels, making it more complicated than just regular hate speech labels.

A deeper investigation reveals that many hate speech datasets in Indonesia contain a minimum amount of labels that are relevant to gender-based hate speech (Ibrahim *et al*., 2022a; Ibrohim *et al*., 2019; Ibrohim and Budi, 2019). The dataset constructed by Ibrohim and Budi (2019) provides

gender-based hate speech although they are outnumbered by the non-hate speech samples at a significant level of imbalance. Using this dataset to develop a gender-based hate speech classification model on Indonesian Twitter posts may cause the model to become overfit and unable to generalize well. Furthermore, collecting new data and manually labeling it could be time and effort-consuming. Therefore, this study attempts to perform data augmentation to increase the number of gender-based hate speech samples while preserving the labels. Previous works have explored data augmentation methods to increase dataset size (Madukwe *et al*., 2022; Azam *et al*., 2022; Cao and Lee, 2020) but mostly on languages other than Indonesian. This study investigates data augmentation methods to classify gender-based hate speech on Indonesian Twitter. Data augmentation is a common strategy for creating new samples that represent the distribution of the original samples while keeping the quality of the augmented data (Madukwe *et al*., 2022). The most extensive study on data augmentation exists in the realm of computer vision due to the straightforward design of simple label-preserving transformations (Bayer *et al*., 2022). It is critical to keep the original labels during data augmentation in text classification to guarantee that the labels are preserved and the changes do not impact the meaning of the data (Madukwe *et al*., 2022).

To the best of the author's knowledge, these challenges have not yet been studied specifically in relation to gender-based hate speech detection. This study attempts to contribute to filling the gaps in the literature by implementing data augmentation methods on the Indonesian Twitter dataset. Therefore, the overall aim of this study is to explore both local and global data augmentation methods. This study also attempts to guarantee that the labels within the augmented data are preserved (Ibrahim *et al*., 2022a). Furthermore, we evaluate the approaches on both local and global data augmentation, as well as the combination of local and global data augmentation, to train the data and develop a gender-based hate speech classification model on the Indonesian Twitter dataset.

Hate Speech

Hate speech is any speech directed at individuals or organizations that contains hateful expressions based on those individuals' or groups' characteristics (Ibrahim *et al*., 2022a). Hatred can be expressed toward someone's race, religion, handicap, sexual orientation, and gender, without being limited to any particular language. Previous work on hate speech detection in the Indonesian language has been conducted by Ibrohim and Budi (2019) where a new dataset was collected and annotated with labels such as hate speech towards race, religion, disability, and gender. Other works construct datasets and develop detection models were investigated by Hendrawan and Al Faraby (2020); Alfina *et al*. (2017); Febriana and Budiarto (2019). In addition, much research in this field also focused on model improvement as in Putri *et al*., (2020); Sevani *et al*. (2021); Taradhita and Darma Putra (2021); Ibrohim *et al*., (2019).

The development of hate speech detection as a text classification problem requires access to a significant volume of clean, unbiased, and balanced datasets (Madukwe *et al*., 2022). It is obvious that the current dataset used in the area of hate speech detection in the Indonesian language suffers from a data scarcity problem as imbalance datasets were used for modeling (Hendrawan and Al Faraby, 2020; Ibrohim and Budi, 2019; Febriana and Budiarto, 2019). Despite duplicating minority class samples as a possible solution to this, data duplication can lead to a variety of problems, including overfitting, poor generalization, and the model being inadequately exposed to various samples from the dataset (Shorten *et al*., 2021). On the other hand, collecting new data comes at the cost of a long process of collection and annotation (Marivate and Sefara, 2020). Therefore, data Augmentation, which involves mixing up the specific forms of language, is one way to address these issues.

Data Augmentation

A common technique for enhancing the quality of existing datasets is data augmentation, which involves creating synthetic samples that closely resemble the distribution of the original samples (Madukwe *et al*., 2022). Researchers work toward a variety of objectives in the domains of data augmentation, including adding more data for dataset classes that are imbalanced or providing extra data for those who have insufficient data. The discipline of computer vision has conducted the most extensive study on data augmentation, in part because simple label-preserving transformations were constructed in an intuitive manner (Bayer *et al*., 2022). An image, for instance, would not change even after being rotated, translated along the x or y axes, or even having the red channel's intensity raised (Shorten *et al*., 2021). Nevertheless, due to the sequential nature of the text and the significance of maintaining the text's semantic and grammatical information, the majority of these methods are not directly applicable to text data (Madukwe *et al*., 2022).

The data augmentation techniques can be separated into two groups: Those that modify just a portion of the original text (local) and those that modify the entire sentence (global) (Madukwe *et al*., 2022). Modifying a portion of the original text (local data augmentation) can be performed by tasks such as replacing words substitution or insertion. Modifying the entire sentence, on the other hand, is accomplished by tasks such as translating the entire statement to another language and then translating back to the original sentence. Other works separate data augmentation techniques based on when they are implemented, such as implementing the

raw data or representation of data (Bayer *et al*., 2022; Marivate and Sefara, 2020). This study focuses on data augmentation techniques based on where they are applied such as in a portion of the original text and the entire sentence (Madukwe *et al*., 2022).

Data augmentation techniques that are applied locally in a portion of the original text is word occur at the word level. An example of this is the insertion, deletion, or substitution of certain words in the document. Using synonyms to substitute words or phrases in data augmentation is the most natural approach (Marivate and Sefara, 2020). A popular approach for this was popularised by Wei and Zou (2019) known as EDA where synonym replacement, random insertion, random swap, and random deletion was constructed. Duong and Nguyen-Thi's (2021) earlier work, each EDA technique is subjected to ten iterations, resulting in 4000 new reviews for every 100 evaluations of the training data. Another similar approach can be performed by replacing a token in the document with is synonym known as token substitution. This can be performed by utilizing language models as in Madukwe *et al*. (2022) where a word in a sentence is masked and BERT attempts to anticipate what the masked word could be by considering the rest of the text.

On the other hand, data augmentation techniques that are applied globally in the entire sentence occur at the document level. A popular example of this is back translation where a word, phrase, sentence, or document is translated into a different language (forward translation) and then back into the original language (back translation) (Bayer *et al*., 2022). The goal of this technique is to construct a new pair of sentences while maintaining the meaning of the source and target sentences (Marivate and Sefara, 2020). In order to construct semantic invariances for the purpose of augmentation, back-translation makes use of the semantic invariances encoded in supervised translation datasets (Shorten *et al*., 2021). Duong and Nguyen-Thi (2021), the Vietnamese source data was translated into English using the Google Translate API and then the augmented data was translated back into Vietnamese.

Materials and Methods

Dataset

A dataset from a previous study (Ibrohim *et al*., 2019) is used where there are tweets with certain categories, such as hate speech labels against race, religion, and gender. Only hate speech labels targeting gender were explored in this study.

Table 1: The characteristics of the selected dataset

	Non-gender hate speech %	Gender-based hate speech %
	12863 (98)	306 (2)
Total	13169 (100)	

As in Table 1, there are 13169 tweets where 306 of them are tweets labeled as gender-based hate speech. Because this dataset is severely imbalanced, a data augmentation step is required. Table 2 shows examples of tweets between gender-based hate speech and non-gender-based hate speech (Ibrahim *et al*., 2022a).

Preprocessing

The dataset contains tweets that are commonly written in an informal style, creating many variations between texts written by one user and another. Some texts and characters in the dataset itself have been modified to protect user confidentiality by transforming the Twitter username to "USER" and any URL link into "URL". Several other preprocessing steps were also performed to standardize the texts. The first preprocessing step is to remove several symbols such as "RT" which indicates that the tweet contains reposted tweet, which is also commonly known as "Re-Tweet". Other symbols that are removed are line breaks, emojis, punctuation, and extra spaces. Another preprocessing step performed is transforming symbols and characters into general ones such as any hashtags into "hashtag" and any number into "number". In order to reduce the inflectional words, word normalization is performed. This process looks at each word and transforms them into its formal word by using a dictionary table in Ibrohim and Budi (2019b). The preprocessing step is conducted in Python, as well as other processes such as data augmentation, feature extraction, and modeling are also conducted in the same programming language. In addition to this, the experiments in data augmenting and modeling use unplug and scikit-learn libraries respectively. Meanwhile, the visualization uses matplotlib and seaborn libraries (Ibrahim *et al*., 2022b; Sagala and Ibrahim, 2022).

Data Augmentation

In this study, both local and global data augmentation methods are explored. Random insertion is a local data augmentation method that randomly inserts synonyms of words in the sentence. On the other hand, back translation is a global data augmentation method that translates the sentence to a certain language and then back to the original language. Lastly, a combination of local and global methods is applied where the back translation is performed first and then followed by random insertion (Ibrahim *et al*., 2022a-b; Arifin *et al*., 2023).

Table 2: Samples of gender-based hate speech

Tweet	Gender-based hate speech
USER Pak USER Mahfud MD sudah berpaling dari Allah SWT Pemberang, demi bangsa ajak semua Buang Islam USER Mr. USER Mahfud MD has turned away from Allah SWT, petulant, for the sake of the nation, he invites all to throw away Islam	No
RT USER: Lihat Kelakuan Trio Cebong Brengsek ini.. Kalau Ketemu Orang2 ini Kalian Mau Apakan twip ? URL RT USER: Look at the behavior of these despicable Trio Cebong. What would you do if you encounter these people, tweeps? URL	
USER Gubernur Indonesia sangat membanggakan. Sangat wibawa sebagai pemimpin. USER The Governor of Indonesia is very admirable. He has a strong leadership presence as a leader	
USER USER Jilbab tapi akhlak bagai pecun\nCuihhh!!!' USER USER Wearing a hijab but behaving like a villain	Yes
USER USER ... Kamu PSK tidak usah nyampur, lepas kerudungmu baru ketauan kamu PSK.... Kamu baca twit nya Fahri yg begitu benci Jokowi itu yg saya balas.... Jokowi itu Presiden tdk pantas dimaki2 oleh seorang Fahri... Twit Fahri penuh	
USER USER … You, a sex worker, don't need to meddle. Take off your hijab and it will be obvious that you're a sex worker…. You read Fahri's tweet that is filled with so much hatred towards Jokowi…. Jokowi is a President and he shouldn't be insulted by someone like Fahri...	
USER USER Makin berkerudung si Transgender ini mkn busuk dan serem kelakuan dan wajahnya! Sundel bolong zaman now\xf0\x9f\xa4\xa3' USER USER The more this transgender person wears a hijab, the fouler and scarier their behavior and face become! Like a Sundel Bolong from nowadays	

In this study, the random insertion utilizes IndoBERT base-uncased to insert a random synonym of a word into the tweet sentences that are written in Indonesian. The inserted words could be any words except predefined stopwords specified in Ibrohim and Budi (2019). In this case, 50% of words in the sentence are augmented, meaning that words' synonyms from half of the sentence will be inserted into the sentence. However, in cases where the sentence is significantly longer than others, only a maximum of 20 synonyms of random words in the sentence will be inserted. This process is described in Algorithm 1.

Algorithm 1: Random Insertion Augmentation (Marivate and Sefara, 2020)

```
words = sentence.split()
aug_max = min(len(words)/2, 20)
num_insert = 0
while num_insert < aug_max:
    get a random word
    if not in stopwords:
        get a synonym of the random word
        insert the synonym randomly
        num_insert += 1
return sentence
```

The back translation method translates tweet sentences into Finnish and then back into Indonesian. The Finnish language was chosen due to its contrast with the original language which is Indonesian. This would provide an alternative to general data augmentation methods as random insertion operates at the word level while back translation modifies the whole sentence. Lastly, another method that is explored is sequential augmentation where backtranslation is performed and then followed by randomly inserting words into the sentences. The back translation algorithm is described in Algorithm 2 (Sagala and Ibrahim, 2022).

Algorithm 2: Back translation augmentation (Marivate and Sefara, 2020)

```
Target_sent = to_target_lang(sentence)
origin_sent = to_origin_lang(sentence)
return origin_sent
```

The augmentation method is performed iteratively until the minority class samples are as many as the majority class samples. This means that gender-based hate speech tweets are repeatedly augmented until both gender-based hate speech and non-hate speech tweets are balanced. However, since the difference between the total number of samples in both classes is significant, the majority class samples are kept at twice as many as the number of samples from the minority class which is 612 samples. This means that the minority class samples are augmented until it reaches 612 samples. This decision is taken by considering that the minority class samples should be augmented at a level where the variation in the augmentation would not reduce the meaning of the samples (Ibrahim *et al*., 2022b).

In order to preserve the label and guarantee that the changes do not impact the meaning of the data, both the original and the augmented are visualized to see if there is an overlap using t-SNE (Madukwe *et al*., 2022). Since the gender-based hate speech class is the only class that is being augmented, this experiment looks at where the original and the augmented data are located. If both original and augmented data are located nearby at the same location, this indicates that the labels are preserved.

Feature Extraction

The next step is to extract features from the dataset, transforming raw text into numerical feature vector representation as input for machine learning algorithms. In this study, the TF-IDF technique is used to extract features. TF-IDF consists of Term-Frequency and Inverse Document Frequency. The Term-Frequency counts the number of the term *t* that appears in document *d* and then is compared to the total number of words in document j. The use of Term-Frequency is critical for presenting the importance of a phrase in a single document. This can be expressed in Eq. 1 (Qi, 2020):

$$TF(i,d) = \frac{Term\ t\ frequency\ in\ document\ d}{Total\ words\ in\ doucment\ d} \quad (1)$$

On the other hand, Inverse Document Frequency looks at the number of documents in the training sample that contain the term *t*, which can be expressed as in Eq. 2 (Qi, 2020):

$$IDF(i) = log_2\left(\frac{Total\ documents}{documents\ with\ term\ d}\right) \quad (2)$$

Inverse document frequency penalizes words that appear in many documents by giving fewer importance values, indicating that these words can be considered noise. The *TF-IDF* incorporates the *TF* and *IDF* values to give more weight to terms that appear frequently in a text but are uncommon across the entire corpus. *TF-IDF* can be expressed as in Eq. 3 (Qi, 2020):

$$TFIDF = TF(i,j) \times IDF(i) \quad (3)$$

TF-IDF is computed for each term in each document, this results in a numerical feature vector representation for the textual data which is used as input for machine learning algorithms. Higher *TF-IDF* values for certain terms represent their importance in document classification. By learning from these feature vectors and the corresponding labels, the machine learning algorithm learns to build a predictive model for unseen documents (Patihullah and Winarko, 2019).

Modeling

The dataset is split into a training set at 80% and a testing set at 20%. Then, a simple model is set as a baseline to compare the performance between models. The baseline model predicts all the test samples as non-hate speech. Machine learning models such as logistic regression, Naïve Bayes, random forest, and XGBoost are used in the experiment. The parameters of these models use default parameters as defined in the sci-kit-learn library. Four distinct algorithms are experimented which were chosen based on prior research (Ibrahim *et al*., 2022b). Logistic regression is a statistical approach used to predict binary categories by analyzing the likelihood of an event occurring (Plaza-Del-Arco *et al*., 2020). This is accomplished by using the logistic function as in Eq. 4:

$$P(y = l) = \frac{1}{1 + exp(-(w.x + b))} \quad (4)$$

The logistic regression formula, as given in Eq. 1, illustrates the prediction process in which the output variable *y* is categorized into a specific label *l* from a collection of labels *L*. Each input word *x* is given a weight *of w* to determine its contribution to the prediction. A random forest is a classification technique made up of many decision trees. Each decision tree in this model evaluates random features. Random Forest's final categorization results are acquired by voting on the judgments of each individual decision tree (Hendrawan and Al Faraby, 2020).

To classify documents, the Naive Bayes classification approach, which is based on the Bayes theorem and makes use of probabilities, is frequently employed. This technique is best suited for classification issues where the features are discrete, such as data on the frequency of words in a document. A document is viewed in this context as a string of words drawn from a particular vocabulary set known as "V" (Plaza-Del-Arco *et al*., 2020). It treats each word independently of the others, assuming that there is no correlation or interaction between them (Andana *et al*., 2019).

The fundamental idea of Extreme Gradient Boosting (XGBoost) is to integrate several separate decision trees, each with a lesser accuracy, into an accurate model. This tree generation procedure employs the gradient descent strategy in which the tree formed in the previous step is built iteratively in the direction of the specified objective function's minimum (Qi, 2020). This iterative method is performed with numerous decision trees until a final prediction model is built with the lowest loss error. The performances of the models are then evaluated and compared by utilizing 5-fold cross-validation. Both accuracy and F-1 scores are used to compare the performance between the models. Overall, the modeling process is described in Algorithm 3.

Algorithm 3: Modeling and evaluation

```
aug_data = [rand_insert, back translate, sequential]
for data in aug_data:
    train_set, test_set = split(data)
    for the model in machine_learning_models:
        model.fit(train_set)
        model.evaluate(test_set)
```

Results and Discussion

Table 3 shows the number of the dataset after being augmented. The total sample size is 1224 for both non-hate speech and gender-based hate speech. The same step is performed for both back translation and random insertion.

The visualization experiment with t-SNE for random insertion, back translation, and sequential augmentation method are shown in Figs. 1-3 respectively. The red dots specify the original gender-based hate tweets while the yellow dots define the augmented gender-based hate tweets. On the other hand, the blue dots show the non-hate tweets. Figures 1-3 show similar patterns where the original and augmented hate tweets are located closely in the same area. This means that the three augmentation methods provide adequate label preservation. However, it also can be seen that there are some non-hate speech tweets and gender-hate speech tweets overlapping with each other. This is caused by some tweets that are difficult to differentiate and could be a case of data annotation issue that leads to ambiguous samples.

Following this, the dataset is split randomly into a training set for 80% and a testing set for 20% which then model training and testing is performed 5 times as in 5-fold cross-validation. Table 4 highlights the accuracy of the dan F1-score of machine learning models trained from the dataset using three data augmentation methods of random insertion, back translation, and sequential combination of both back translation and random insertion.

The baseline model in Table 4 obtains 53 and 35% for accuracy and F1-score respectively. In terms of data augmentation methods, random insertion provides slightly better performance in both accuracy and F1-score compared to back translation and sequential combination methods. This means that randomly inserting synonyms of words into the tweet sentences provides better data augmentation for the classification task in this case. Both back translation and sequential methods generally lead to a high number of accuracy and F1-score values. However, none of the model's performance outnumbers the model's performance of models trained with a dataset generated from the random insertion method. This means that translating tweet sentences into Finnish and then back into Indonesian would not provide better data augmentation in this case. Furthermore, the model trained with the original dataset achieves slightly lower accuracy and F1-score compared to models trained with the dataset generated using data augmentation methods. Overall, data augmentation improves the quality of data as both evaluation parameters obtain better performance for models trained with datasets generated from data augmentation.

Table 3: The distribution of gender and non-gender-based hate speech samples after performing back translation and random insertion

		Hate speech_gender	
	Non-hate speech	---	
	Original	Original	Augmented
	612	306	306
Total	612	612	

Table 4: Accuracy and F1-score value for each machine learning model trained with dataset generated from augmentation methods

		No augmentation	Random insertion	Backtranslation	Sequential
Accuracy	Baseline	0,652	0,534	0,534	0,534
	Logistic regression	0,926	0,970	0,927	0,924
	Naïve Bayes	0,923	0,944	0,931	0,924
	Random forest	0,947	0,971	0,934	0,933
	XGBoost	0,934	0,960	0,902	0,892
F1-score	Baseline	0,394	0,348	0,348	0,348
	Logistic regression	0,914	0,970	0,927	0,924
	Naïve Bayes	0,915	0,944	0,931	0,924
	Random forest	0,938	0,971	0,934	0,933
	XGBoost	0,924	0,960	0,902	0,89

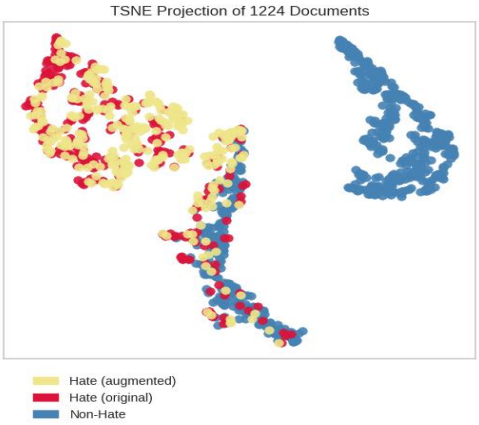

Fig. 1: The graph shows the t-SNE-based visualization of the gender-hate speech tweets between the original and the augmented data using the random insertion method

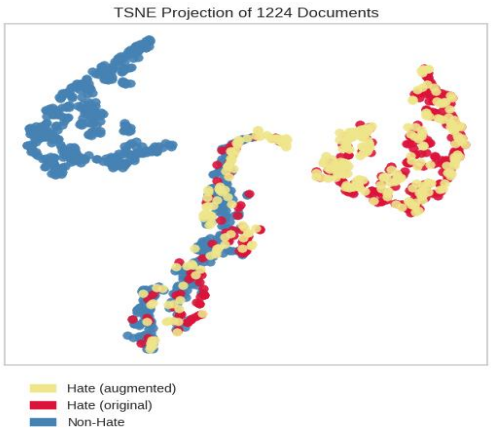

Fig. 2: The graph shows the t-SNE-based visualization of the gender-based hate speech tweets between the original and the tweets generated using back translated data augmentation method

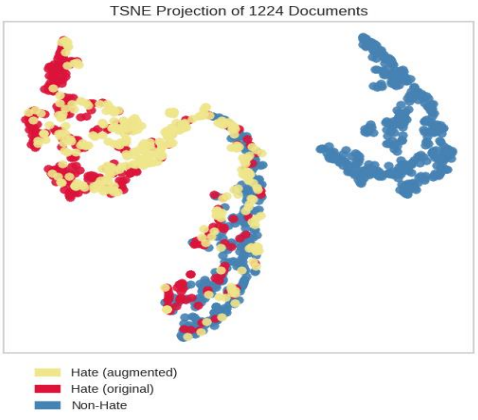

Fig. 3: The graph shows the t-SNE-based visualization of the gender-based hate speech tweets between the original and augmented data using a combination of back translation and random insertion in the gender-hate speech class

Conclusion

In conclusion, this study attempts to classify non-hate speech tweets and gender-based hate speech tweets in the Indonesian language by exploring data augmentation methods to tackle datasets with class imbalance issues. Several data augmentation methods are experimented such as random insertion, back translation, and sequential combination of both random insertion and back translation. Several machine learning models are used to develop classification models to classify gender-based hate speech and non-hate speech tweets. The experiment shows that models trained with a dataset generated from the random insertion data augmentation method lead to higher accuracy and F1 score compared to other data augmentation methods. This demonstrates that augmenting the dataset at the word level by inserting random synonyms of words into the sentence could improve the quality of the dataset. In comparison, globally augmenting the dataset at the sentence level, in this case, does not significantly enhance the quality of the dataset as well as locally augmenting the dataset at the sentence level. Despite slight differences in the model's performance, the experimented data augmentation methods are able to preserve the labels as confirmed by the t-SNE visualization results. The benefit of this study is that data augmentation could provide solutions for highly imbalanced datasets by increasing the amount of the minority class while preserving the labels. In the future, there are some opportunities to improve this study. For instance, exploring other data augmentation methods such as substitution and the combination between insertion, substitution, and translation. Other languages could also be utilized for performing translation in the back translation method.

Acknowledgment

The authors of this study would like to express their sincere gratitude and appreciation to the experts who provided valuable input to improve the quality of this study. The researchers would also like to thank Bina Nusantara University for the support they provided with the funding and infrastructure facilities.

Funding Information

This project has received financial support from Bina Nusantara. The financial support is financed by the office of research and technology transfer.

Author's Contributions

Muhammad Amien Ibrahim: Coding, written and finished the manuscript.

Samsul Arifin: Edited, implemented and finished the manuscripts.

Eko Setyo Purwanto: Organizing the basis for the theory and the implemented methods.

Ethics

This article's material is original and has not been published before. The corresponding author confirms that this study has no conflicts of interest or ethical problems.

References

Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017, October). Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 233-238). IEEE. https://doi.org/ 10.1109/ICACSIS.2017.8355039

Amjad, M., Ashraf, N., Zhila, A., Sidorov, G., Zubiaga, A., & Gelbukh, A. (2021). Threatening language detection and target identification in Urdu tweets. *IEEE Access*, *9*, 128302-128313. https://doi.org/10.1109/ACCESS.2021.3112500

Andana, E. K., Othman, M., & Ibrahim, R. (2019). Comparative analysis of text classification using Naive Bayes and Support Vector Machine in detecting negative content in Indonesian twitter. *Comparative Analysis of Text Classification Using Naive Bayes and Support Vector Machine in Detecting Negative Content in Indonesian Twitter*, *8*(1.3), 356-362. https://doi.org/10.30534/ijatcse/2019/6481.32019

Arifin, S., Nicholas, A., Baskoroputro, H., Prabowo, A. S., Ibrahim, M. A., & Rahayu, A. (2023). Algorithm for Digital Image Encryption Using Multiple Hill Ciphers, a Unimodular Matrix and a Logistic Map. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(6s), 311-324. https://ijisae.org/index.php/IJISAE/article/view/2858

Azam, U., Rizwan, H., & Karim, A. (2022, June). Exploring Data Augmentation Strategies for Hate Speech Detection in Roman Urdu. In *Proceedings of the 13th Language Resources and Evaluation Conference*, (pp. 4523-4531). https://aclanthology.org/2022.lrec-1.481

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, (pp. 54-63). https://doi.org/10.18653/v1/s19-2007

Bayer, M., Kaufhold, M. A., & Reuter, C. (2022). A survey on data augmentation for text classification. *ACM Computing Surveys*, *55*(7), 1-39. https://doi.org/10.1145/3544558

Cao, R., & Lee, R. K. W. (2020, December). Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6327-6338). https://doi.org/10.18653/v1/2020.coling-main.557

Duong, H. T., & Nguyen-Thi, T. A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1), 1-16. https://doi.org/10.1186/s40649-020-00080-x

Febriana, T., & Budiarto, A. (2019, August). Twitter dataset for hate speech and cyberbullying detection in Indonesian language. In *2019 International Conference on Information Management and Technology (ICIMTech)*, (Vol. *1*, pp. 379-382). IEEE. https://doi.org/ 10.1109/ICIMTech.2019.8843722

García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., & Valencia-García, R. (2023). Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, *9*(3), 2893-2914. https://doi.org/10.1007/s40747-022-00693-x

Hendrawan, R., & Al Faraby, S. (2020, August). Multilabel classification of hate speech and abusive words on Indonesian Twitter social media. In *2020 International Conference on Data Science and Its Applications (ICoDSA)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICoDSA50139.2020.9212962

Ibrahim, M. A., Sagala, N. T. M., Arifin, S., Nariswari, R., Murnaka, N. P., & Prasetyo, P. W. (2022a, July). Separating Hate Speech from Abusive Language on Indonesian Twitter. In *2022 International Conference on Data Science and Its Applications (ICoDSA)* (pp. 187-191). IEEE. https://doi.org/10.1109/icodsa55874.2022.9862850

Ibrahim, M. A., Arifin, S., Yudistira, I. G. A. A., Nariswari, R., Abdillah, A. A., Murnaka, N. P., & Prasetyo, P. W. (2022b). An Explainable AI Model for Hate Speech Detection on Indonesian Twitter. *CommIT (Communication and Information Technology) Journal*, *16*(2), 175-182. https://doi.org/10.21512/commit.v16i2.8343

Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the 3rd Workshop on Abusive Language Online* (pp. 46-57). https://doi.org/10.18653/v1/W19-3506

Ibrohim, M. O., Setiadi, M. A., & Budi, I. (2019, November). Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features. In *Proceedings of the 1st International Conference on Advanced Information Science and System*, (pp. 1-5). https://doi.org/10.1145/3373477.3373495

Lin, K. Y., Lee, R. K. W., Gao, W., & Peng, W. C. (2021, December). Early prediction of hate speech propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*, (pp. 967-974). IEEE. https://doi.org/10.1109/ICDMW53433.2021.00126

Madukwe, K. J., Gao, X., & Xue, B. (2022). Token replacement-based data augmentation methods for hate speech detection. *World Wide Web*, *25*(3), 1129-1150. https://doi.org/10.1007/s11280-022-01025-2

Marivate, V., & Sefara, T. (2020). Improving short text classification through global augmentation methods. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25-28, 2020, Proceedings 4*, (pp. 385-399). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_21

Meng, Q., Suresh, T., Lee, R. K. W., & Chakraborty, T. (2023). Predicting hate intensity of twitter conversation threads. *Knowledge-Based Systems*, 110644. https://doi.org/10.1016/j.knosys.2023.110644

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS One*, *15*(8), e0237861. https://doi.org/10.1371/journal.pone.0237861

Patihullah, J., & Winarko, E. (2019). Hate speech detection for indonesia tweets using word embedding and gated recurrent unit. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(1), 43-52. https://doi.org/10.22146/ijccs.40125

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors*, *19*(21), 4654. https://doi.org/10.3390/s19214654

Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120. https://doi.org/10.1016/j.eswa.2020.114120

Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2020). Detecting misogyny and xenophobia in Spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, *20*(2), 1-19. https://doi.org/10.1145/3369869

Putri, T. T. A. Sriadhi, S. Sari, R. D. Rahmadani, R. & Hutahaean, H. D. (2020). A comparison of classification algorithms for hate speech detection, *IOP Conf. Ser. Mater. Sci. Eng.*, Vol. *830*, no. 3 https://doi.org/10.1088/1757-899X/830/3/032006

Qi, Z. (2020, June). The text classification of theft crime based on TF-IDF and XGBoost model. In *2020 IEEE International conference on artificial intelligence and computer applications (ICAICA)* (pp. 1241-1246). IEEE. https://doi.org/10.1109/ICAICA50127.2020.9182555.

Sagala, N. T., & Ibrahim, M. A. (2022, July). A Comparative Study of Different Boosting Algorithms for Predicting Olympic Medal. In *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICCED56140.2022.10010351

Sevani, N., Soenandi, I. A., & Wijaya, J. (2021, October). Detection of Hate Speech by Employing Support Vector Machine with Word2Vec Model. In *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICEEIE52663.2021.9616721

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, *8*, 1-34. https://doi.org/10.1186/s40537-021-00492-0

Taradhita, D. A. N., & Darma Putra, I. (2021). Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network. *Journal of ICT Research & Applications*, *14*(3). https://doi.org/ 10.5614/itbj.ict.res.appl.2021.14.3.2.

Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*. https://doi.org/10.48550/arXiv.1901.11196