Journal of Computer Science

Original Research Paper

Prompt-Based Data Augmentation with Large Language Models for Indonesian Gender-Based Hate Speech Detection

[1]Muhammad Amien Ibrahim, [2]Faisal, [1]Zefanya Delvin Sulistiya and [1]Tora Sangputra Yopie Winarto

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

Article history
Received: 05-03-2024
Revised: 15-04-2024
Accepted: 22-04-2024

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: The increasing amount of content on social media content makes the use of automatic moderation crucial for preserving a healthy online community and reducing the spread of offensive and abusive content, such as hate speech based on gender. Developing automated social media moderation using machine learning demands a large and balanced dataset. However, difficulties such as data scarcity and class imbalance have hindered the development of gender-based hate speech detection on Indonesian Twitter communities. Creating and annotating a new dataset would be time-consuming and costly. One practical alternative is to use data augmentation methods to help address the minority class imbalance in datasets. This study investigates how prompt-based data augmentation may be used with a large language model to provide organic tweet samples for gender-based hate speech detection. Furthermore, the study investigates the preservation of labels in augmented Twitter samples. In comparison to the benchmark back translation approach, the results show that prompt-based data augmentation using a large language model may generate new and organic Twitter samples while keeping labels preserved and avoiding memorization. In conventional machine learning models, prompt-based data augmentation with a large language model shows competitive performance compared to back translation in terms of accuracy metrics. According to these results, using prompting for data augmentation on large language models is an alternative strategy that can provide new, less memorization tweet samples that maintain label integrity while achieving competitive accuracy results.

Keywords: Hate Speech Detection, Data Augmentation, Large Language Models

Introduction

The use of social media as a communication tool has become critical in today's culture. However, social media is frequently used for disseminating hate speech. Hate speech on social media platforms such as Twitter can range from politics, social concerns, sports, religion and gender. Furthermore, hate speech directed towards a particular gender can fuel prejudice, discrimination and the ongoing spread of negative perceptions. Online hate speech against women has major real-world consequences, including legal action against social media platforms for inadequate moderation of offensive posts (Frenda *et al*., 2019).

Given the widespread distribution of hate speech towards gender on social media platforms such as X (formerly known as Twitter), it is critical to use social media moderation to identify and prevent the spread of such content. However, it would take an extensive amount of time and resources to manually monitor social media conversations. Machine learning techniques such as text classification offer an alternative to automatic hate speech detection. In order for the machine learning model to recognize patterns and identify hate speech on social media, these techniques require a significant amount of data.

In the case of Ibrohim and Budi (2019), efforts were made to compile a dataset with more specific labels for the purpose of detecting hate speech from Indonesian Twitter communities. This dataset has been crucial to several earlier research projects in this field (Elisabeth *et al*., 2020; Marpaung *et al*., 2021; Ibrohim *et al*., 2020). Nonetheless, one of the challenges to more research in this area is the absence of a sufficient dataset for a more granular categorization, such as a gender-based hate speech dataset.

The gender-based hate speech dataset (Ibrohim and Budi, 2019) shows a significant class imbalance. This imbalance is likely to result in the gender-based hate speech detection model that is prone to overfitting. While gathering more data is a possible solution, the procedure would be costly in terms of time and annotation resources. In this study, an effort is made to increase the number of the dataset by using data augmentation techniques.

Data augmentation is a popular method for improving the quality of already-existing datasets, which involves making synthetic samples that closely resemble the original samples' distribution (Madukwe *et al*., 2022). Many studies on data augmentation are conducted in the field of computer vision. This is due to the natural use of simple, label-preserving transformations to images (Bayer *et al*., 2022). This includes rotation, translation, color intensification and other operations of a similar nature (Shorten *et al*., 2021). Unfortunately, most of these approaches do not translate well to text due to the sequential structure included in textual data. This is because maintaining semantic and syntactic information inside the textual environment is crucial (Madukwe *et al*., 2022).

As regards text data, one technique to implement data augmentation is local data augmentation. This technique consists of substitution and insertion of words inside the document where some words are altered to produce artificial samples. Alternatively, another technique that can be used is global data augmentation where the entire document can be altered to produce artificial samples. For instance, a full paragraph can be translated into one language and then back into the original language.

Current studies demonstrate the effectiveness of a prompting-based approach for labeled training data generation using Large Language Models (LLMs). This technique supports data augmentation by reformulating each sentence in the training samples into several instances that are conceptually similar but semantically diverse (Dai *et al*., 2023).

Previous research has looked into both local and global data augmentation techniques for gender-based hate speech detection (Madukwe *et al*., 2022; Azam *et al*., 2022; Ibrahim *et al*., 2023). However, there has been little investigation into data augmentation via a prompting-based strategy using large language models. As a result, the goal of this study is to build synthetic samples using large language models as a data augmentation approach to gender-based hate speech identification in Indonesian Twitter.

Hate Speech

Hate speech, as described in Ibrohim and Budi (2019), is any verbal abuse or threats that express prejudice based on characteristics they possess. Without being limited to any one language, this might include characteristics such as religion, gender and race. The identification of hate speech or similar related areas has been extensively studied in the past in a variety of languages (Azam *et al*., 2022; Leite *et al*., 2020; Hendrawan *et al*., 2020; Sutejo and Lestari, 2018; Perifanos and Goutsos, 2021). Notably, Ibrohim and Budi (2019) pioneered multi-label hate speech identification in the Indonesian language, developing and testing a new dataset. This study effectively annotated multiple tweets using criteria such as hate speech level, race/ethnicity and gender/sexual orientation, laying the groundwork for hate speech identification in the Indonesian language.

Many developments in research on hate speech detection in the Indonesian language have focused on comparing classification model performance across various machine learning models and architectures, as seen in Marpaung *et al*. (2021); Kurniawan and Budi (2020); Taradhita *et al*. (2021); Pratiwi *et al*. (2018); Briliani *et al*. (2019); Erizal *et al*. (2019). Alternatively, other research has changed its emphasis to comparing model performance on the basis of different features, such as lexicons, word embeddings, TF-IDF, unigrams and Part of Speech tags (Ibrohim *et al*., 2020; Aulia and Budi, 2019).

Hate speech detection in Indonesia has mostly focused on binary categorization of hate speech against non-hate speech. Nonetheless, a significant amount of research has been done in the past on more granular category classification of hate speech in the English language. Among this research are the classification of sexism (Rodriguez Sanchez *et al*., 2020), the detection of misogyny (Pamungkas *et al*., 2020) and the identification of online hate speech directed against women (Frenda *et al*., 2019). To the knowledge of the author, the closest work on hate speech recognition in Indonesian that delves into more granular categories is the study on abusive language detection presented by Ibrahim *et al*. (2022). The development of a more specific hate speech category classification in the Indonesian language may be hindered by an inadequate amount of samples or a severe imbalance within the category in the dataset, in Ibrohim and Budi (2019).

Data Augmentation

One solution for tackling data imbalance issues is employing data augmentation. Data augmentation is an established method for increasing the number of samples in the existing datasets. It involves creating synthetic samples that closely resemble the original samples' distribution (Madukwe *et al*., 2022). In the field of computer vision, much effort has been focused on improving datasets through the use of data augmentation techniques. This entails methodically altering pictures while maintaining the essential semantic meaning (Bayer *et al*., 2022). With a generic image, properties are preserved even after operations such as rotation, translation along the axes, or modifications to particular color channels (Shorten *et al*., 2021). However, the application of such operations to text data provides challenges due to the sequential pattern of textual information. Many typical approaches used in computer vision, which show success with picture data, may not be immediately applicable to text. As a result, particular and careful techniques must be taken throughout augmentation operations to preserve the text's semantic and syntactic coherence, ensuring that the changes do not compromise the underlying meaning of the sentences (Madukwe *et al.*, 2022).

In general, data augmentation techniques for text data are divided into two categories based on how and where they are applied to the text: Local and global augmentation (Madukwe *et al.*, 2022). Local augmentation modifies sentences at the word level. This may be accomplished by inserting or removing random words from the sentence. Moreover, a sequence of word deletion and insertion can also be used for the local augmentation technique, often referred to as the substitution operation. The substitution operation is the most natural way for the local augmentation method (Marivate and Sefara, 2020). Wei and Zou (2019) investigate a popular approach for local augmentation known as EDA, which includes multiple local augmentation techniques such as synonym replacement, random insertion, random swap and random deletion. Another technique for local augmentation uses language models such as BERT to replace masked words with the most likely tokens based on the context of the sentences (Madukwe *et al.*, 2022).

In contrast, global augmentation operates at the document level. Back translation is a popular approach for global augmentation, which involves translating a text into another language and then translating it back into the original language (Bayer *et al.*, 2022). This technique creates a new document sample that retains the original meaning (Marivate and Sefara, 2020). Back translation for data augmentation has been used in earlier research, using a variety of languages (Ibrahim *et al.*, 2023; Duong and Nguyen-Thi, 2021; Kumar *et al.*, 2021).

Some NLP researchers have shifted their attention to developing advanced techniques to enhance the data augmentation approach through large language models such as GPT-3. One such technique is GPT3Mix, in which Yoo *et al*. (2021) demonstrate a way to create artificial text samples from a combination of real samples (Yoo *et al.*, 2021). Using a small sample size of sentences chosen from the task-specific training data, GPT3Mix combines these samples into the prompt to create an augmented sentence that is influenced by the sample sentences (Yoo *et al.*, 2021). In a similar direction, (Sahu *et al.*, 2022) propose applying a prompting-based strategy to create labeled training data for intent classification using GPT3, hence improving the performance of intent classifiers, particularly when training data is limited (Sahu *et al.*, 2022). In addition to this, (Dai *et al.*, 2023) introduced AugGPT, a text data augmentation strategy based on ChatGPT that rephrases each sentence in the training samples into similar but semantically distinct samples for use in downstream model training.

Materials and Methods

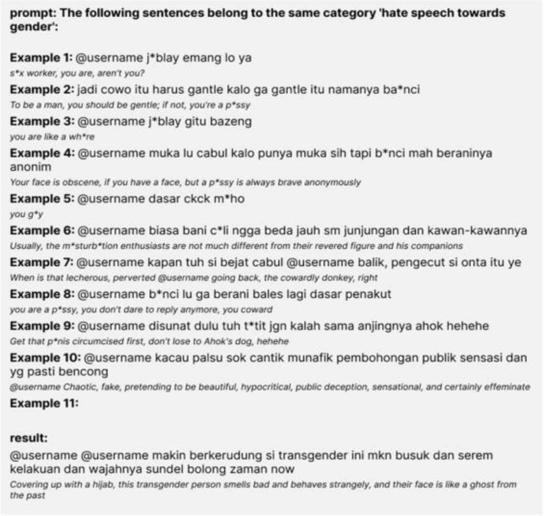
Dataset and Preprocessing

Using a dataset from prior work (Ibrohim and Budi, 2019), this study focused on analyzing hate speech labels targeting gender. The dataset includes tweets labeled with hate speech against race, religion and gender. The dataset includes tweets written in an informal and varied communication style, using abbreviations and slang phrases. The variety in communication style is due to differences in people's personalities, personal preferences, cultural backgrounds and education, which result in significant discrepancies in messages posted by various users on Twitter. As a result, a preprocessing step is implemented to standardize the text through text substitutions. For instance, all Twitter usernames in the dataset are replaced with "USER," and all numerical values are replaced with "9999." Additional text substitution is used, such as replacing hashtags with the "#hashtag." Furthermore, text cleaning techniques were used, such as the removal of the "RT" symbol that represents Re-Tweet, as well as the removal of emojis, punctuation marks and excessive spaces. To reduce the effect of inflectional variations, a normalization operation is performed. This entails going over each term and utilizing a dictionary table in Ibrohim and Budi (2019) to convert it into its formal equivalent.

Data Augmentation

Due to the dataset's significant imbalance, which includes a total of 13,169 tweets with just 306 categorized as gender-based hate speech, data augmentation is an essential step. This study investigates two methods of augmenting data: Back translation and sample generation using a large language model. Back translation is a technique for global data augmentation in which a sentence is translated into a target language and then back into the original language (Duong and Nguyen-Thi, 2021). Back translation is chosen for its ability to enhance and change complete samples globally, similar to the second approach for creating samples using a large language model. It is assumed that using a large language model to augment data allows for the global generation of new samples, which results in the creation of whole new sentences.

Back translation is the process of transforming text samples into a foreign language and then translating them back into their original language. This approach is language-independent and the Helsinki-NLP model is used in this study to translate Indonesian tweets into Finnish and then back into Indonesian. Finnish was selected specifically for exploration purposes.



prompt: The following sentences belong to the same category 'hate speech towards gender':

Example 1: @username j*blay emang lo ya
*s*x worker, you are, aren't you?*

Example 2: jadi cowo itu harus gantle kalo ga gantle itu namanya ba*nci
*To be a man, you should be gentle; if not, you're a p*ssy*

Example 3: @username j*blay gitu bazeng
*you are like a wh*re*

Example 4: @username muka lu cabul kalo punya muka sih tapi b*nci mah beraninya anonim
*Your face is obscene, if you have a face, but a p*ssy is always brave anonymously*

Example 5: @username dasar ckck m*ho
*you g*y*

Example 6: @username biasa bani c*li ngga beda jauh sm junjungan dan kawan-kawannya
*Usually, the m*sturb*tion enthusiasts are not much different from their revered figure and his companions*

Example 7: @username kapan tuh si bejat cabul @username balik, pengecut si onta itu ye
When is that lecherous, perverted @username going back, the cowardly donkey, right

Example 8: @username b*nci lu ga berani bales lagi dasar penakut
*you are a p*ssy, you don't dare to reply anymore, you coward*

Example 9: @username disunat dulu tuh t*tit jgn kalah sama anjingnya ahok hehehe
*Get that p*nis circumcised first, don't lose to Ahok's dog, hehehe*

Example 10: @username kacau palsu sok cantik munafik pembohongan publik sensasi dan yg pasti bencong
@username Chaotic, fake, pretending to be beautiful, hypocritical, public deception, sensational, and certainly effeminate

Example 11:

result:

@username @username makin berkerudung si transgender ini mkn busuk dan serem kelakuan dan wajahnya sundel bolong zaman now
Covering up with a hijab, this transgender person smells bad and behaves strangely, and their face is like a ghost from the past

Fig. 1: Sample prompt and its corresponding result

Each gender-based hate speech tweet is subjected to iterative back-translation augmentation. With 306 original gender-based hate speech tweets present, each of these tweets is back-translated once, resulting in a new set of 306 enhanced gender-based hate speech tweets. As a result, the dataset contains 612 gender-based hate speech tweets, half of which are original and the other half are augmented. In order to preserve variance in the augmented sample without compromising the original meaning of the samples, back translation is only performed once.

In contrast, a large language model generates an equal amount of augmented gender-based hate speech tweets. The augmentation procedure involves prompting a GPT3 engine with a template prompt adapted from (Sahu *et al.*, 2022), in Fig 1. Using ten random examples of gender based hate speech tweets, GPT-3 completes the 11th sample, effectively generating a new instance of a gender based hate speech tweet. The resulting text's variety in GPT-3 is determined by the top-p and temperature parameters. A larger value of the top-p parameter explores more alternative words, boosting variety, whereas a smaller value restricts the selection, limiting diversity by prioritizing tokens with greater probability. Another GPT3 parameter is temperature, where a higher value leads to more creative and less predictable outputs by increasing the possibility of less probable tokens and decreasing the likelihood of more probable ones. This prompting technique is repeated 306 times, which corresponds to the total number of original gender-based hate speech tweets, resulting in 612 examples of the label indicated above. Half of these 612 examples are the original tweets, while the other half are augmented samples.

Both the original and augmented samples are visualized using t-SNE to verify that augmented samples are preserved and that the two investigated augmentation techniques don't affect the meaning of the tweets. Given that the only tweets being augmented are gender-based hate speech tweets, this experiment focuses on assessing the closeness of the original and augmented tweet samples. The presence of both the original and augmented data in the same area indicates that the labels have been properly preserved.

To input raw text into machine learning models, the subsequent step is feature extraction from the dataset, which involves transforming raw text into numerical representations. The TF-IDF approach, which combines Term-Frequency (TF) and Inverse Document Frequency (IDF), is utilized in this study to extract features.

In order to evaluate the performance of two augmented datasets derived from different augmentation techniques for gender-based hate speech classification, a 5-fold cross-validation approach is utilized. Several conventional machine learning classification models, including Logistic Regression, Naïve Bayes, Random Forest and XGBoost, are employed for evaluation. The parameters for these models are determined by the default parameters specified in the scikit-learn package. These models are selected to evaluate performance since simpler models are adequate to complete the task at hand. The accuracy metric is used to compare performance across different models and augmentation techniques.

Results and Discussion

Table 1 summarizes the complete dataset, including both original and augmented tweet samples. There are 612 randomly selected non-hate speech tweets and an equal number of gender-based hate speech tweets in total. Of the tweets in the latter group, 306 are original while the other 306 are augmented. It is critical to note that the 306 augmented tweet samples have two versions: One obtained from the back translation technique and one generated by the large language model.

The effects of the GPT-3 temperature parameter on data augmentation accuracy are investigated. Figure 2, lower temperature parameter values result in improved accuracy across the four models explored. This shows that less diverse and innovative words lead to higher accuracy levels. Similarly, variations in the top-p parameter affect the accuracy of the four models explored in Fig. 3. Accuracy improves with smaller top-p values and eventually decreases as the top-p parameter increases. This suggests that when GPT-3 samples a smaller selection of words, the generated text is less diversified, resulting in greater accuracy results.

Figures 2-3, the highest accuracy across the four models is achieved with top-p and temperature parameter values of 0.4 and 0.25, respectively. A GPT-3 model with these parameter values is used to create a set of samples. The resulting accuracy is then compared to the back translation technique, in Fig. 6 The results show that the large language model generated samples perform competitively to the back translation technique.

Table 1: Summary of the augmented dataset using two augmentation techniques

	Non-hate speech	Gender-based hate speech	
	Original	Original	Back-translated augmentation
	612	306	306
Total set 1	1224		
	Original	Original	Prompt-based LLM augmentation
	612	306	306
Total set 2	1224		

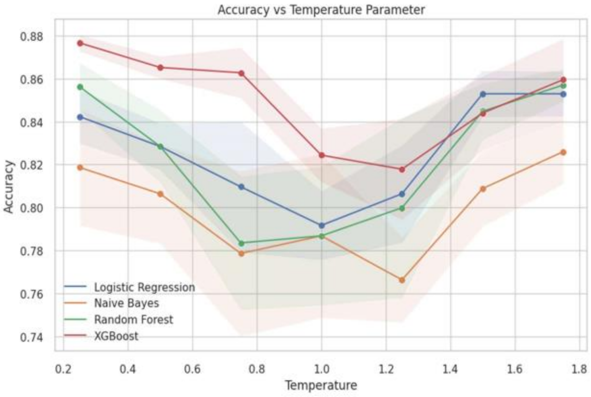


Fig. 2: Influence of temperature parameter on model accuracy

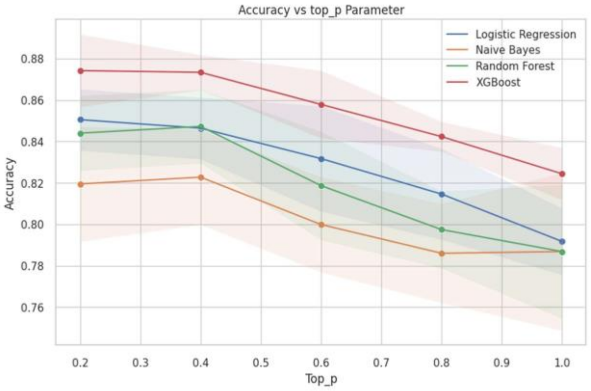


Fig. 3: Influence of temperature parameter on model accuracy

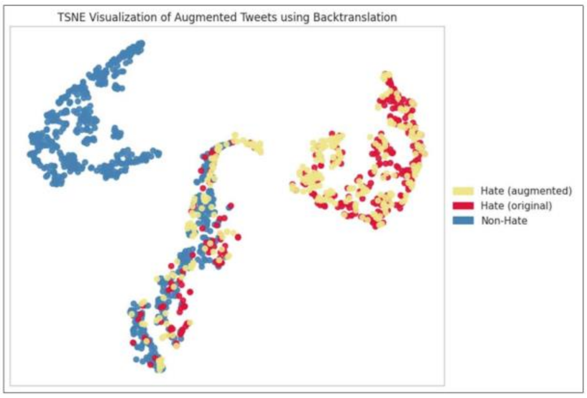


Fig. 4: Back translation augmentation: Original vs. augmented data

Figures 4-5 show the visualization experiments that used t-SNE for both the back translation augmentation technique and the large language model generated technique. The red dots indicate original gender-based hate tweets, the yellow dots indicate augmented gender-based hate tweets and the blue dots indicate non-hate tweets.

On the right side, Fig. 4 shows that back translation is able to produce gender-based hate speech tweet samples that resemble the original gender-based hate speech tweet samples indicating that both groups are located in a close area. This means that data augmentation using back translation is able to generate new data samples with similar variance to the original gender-based hate speech tweet samples. However, this finding might indicate that memorizing is occurring, where specific examples are being augmented to the tweet samples rather than understanding their underlying patterns, which could lead to overfitting. It is interesting to note that the middle area in Fig. 4 shows that there are three labels surrounding one area. This is likely because there are some similar tweets where some of them are labeled as non-hate while others are labeled as gender-based hate speech, meaning that mislabelling is a possible cause for this. Thus, back-translated samples are visible in this area since back-translation augments the same tweets.

The top right area of Fig. 5 shows how Large Language Model gender-based hate speech generated tweet samples are located which is closely resembling the original gender-based hate speech tweet samples. Similar to Fig. 4, there are areas where both non-hate and gender-based hate speech tweets are in the same area, indicating a possible mislabeling. However, one interesting thing to note from Large Language-generated samples is that the augmented tweet sample is not present in this area. Instead, they are located close to the area of original gender-based hate speech tweet samples and even some of them are in different areas but still close to the original gender-based hate speech tweet samples. This means that data augmentation using the Large Language Model is able to generate new variance of samples while maintaining the original meaning of the tweet samples as indicated with competitive accuracy performance in Fig. 6.

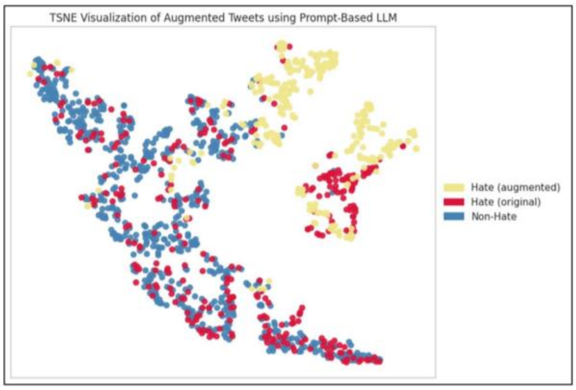


Fig. 5: Prompt-based LLM augmentation: Original vs. augmented data

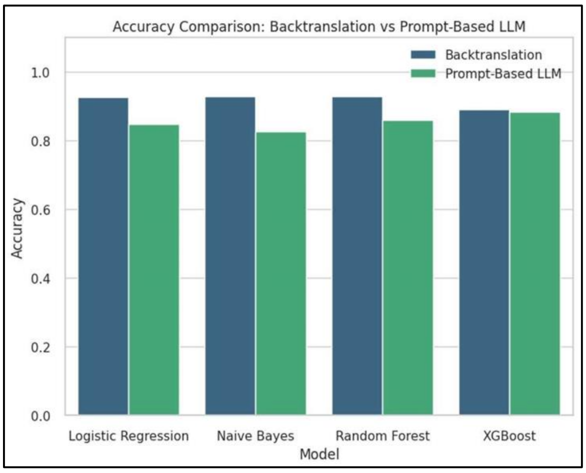


Fig. 6: Model performance comparison between the benchmark back translation approach vs. prompt-based

Conclusion

Finally, this study investigates data augmentation with a large language model to address imbalances in a dataset of gender-based hate speech tweets. Top-up and temperature parameter values of 0.4 and 0.25 provide optimal results for gender-based hate speech classification in conventional models such as Logistic Regression, Naïve Bayes, random forest and XGBoost. While achieving comparable accuracy to back translation data augmentation, the large language model-generated data augmentation technique generates samples that are distinct from the original tweet samples. This suggests that the large language model technique has the ability to produce higher-quality samples than back translation while retaining competitive accuracy in gender-based hate speech classification.

In the future, further studies in this area might focus on the impact of repeated augmentation of the same original tweet sample with the large language model-generated data augmentation technique, compared to back translation. As discussed, a back translation could produce similar augmented tweets with lower variation, making it vulnerable to overfitting due to the large variation in the augmented tweets, as opposed to Large Language Model-generated tweets. As a result, analyzing the sustainability and long-term consequences of these techniques is critical for gaining a thorough knowledge of their effectiveness.

Acknowledgment

The researchers extend their sincere thanks and deep appreciation to the knowledgeable professionals who offered insightful feedback to enhance the caliber of this research endeavor. Additionally, they convey their gratitude to Bina Nusantara University for the vital support provided through funding and access to necessary facilities and infrastructure that enabled the successful completion of this study.

Funding Information

This study is supported by Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's Associate Research Grant entitled Prompt-Based Data Augmentation with Large Language Models for Indonesian Gender-Based Hate Speech Detection with contract number: 069A/VRRTT/III/2024 and contract date: March 18, 2024.

Author's Contributions

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

Faisal: Conceptualization, methodology and finished the manuscripts.

Zefanya Delvin Sulistiya and Tora Sangputra Yopie Winarto: Experimentation and finished the manuscripts.

Ethics

This article contains original material that has not previously been published. The corresponding author claims that this study is free of conflicts of interest or ethical issues.

References

Aulia, N., & Budi, I. (2019). Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach. *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, 164-169. https://doi.org/10.1145/3330482.3330491

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

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

Briliani, A., Irawan, B., & Setianingsih, C. (2019). Hate Speech Detection in Indonesian Language on Instagram Comment Section Using K-Nearest Neighbor Classification Method. *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 98-104. https://doi.org/10.1109/iotais47347.2019.8980398

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2023). AugGPT: Leveraging ChatGPT for Text Data Augmentation. *ArXiv*, 2302.13007. https://doi.org/10.48550/arXiv.2302.13007

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

Elisabeth, D., Budi, I., & Ibrohim, M. O. (2020). Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study. *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 1-6. https://doi.org/10.1109/icoict49345.2020.9166251

Erizal, E., Irawan, B., & Setianingsih, C. (2019). Hate Speech Detection in Indonesian Language on Instagram Comment Section Using Maximum Entropy Classification Method. *2019 International Conference on Information and Communications Technology (ICOIACT)*, 533-538. https://doi.org/10.1109/icoiact46704.2019.8938593

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online Hate Speech against Women: Automatic Identification of Misogyny and Sexism on Twitter. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4743-4752. https://doi.org/10.3233/jifs-179023

Ibrahim, M. A., Arifin, S., & Purwanto, E. S. (2023). Exploring Data Augmentation for Gender-Based Hate Speech Detection. *Journal of Computer Science*, *19*(10), 1222-1230. https://doi.org/10.3844/jcssp.2023.1222.1230

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

Ibrohim, M. O., & Budi, I. (2019). Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46-57. https://doi.org/10.18653/v1/w19-3506

Ibrohim, M. O., Setiadi, M. A., & Budi, I. (2020). Identification of hate speech and abusive language on indonesian Twitter using the Word2vec, part of speech and emoji features. *Proceedings of the 1st International Conference on Advanced Information Science and System*, 1-5. https://doi.org/10.1145/3373477.3373495

Kumar, V., Choudhary, A., & Cho, E. (2021). Data Augmentation using Pre-trained Transformer Models. *ArXiv*, 2003.02245. https://doi.org/10.48550/arXiv.2003.02245

Kurniawan, S., & Budi, I. (2020). Indonesian Tweets Hate Speech Target Classification using Machine Learning. *2020 5th International Conference on Informatics and Computing (ICIC)*, 1-5. https://doi.org/10.1109/icic50835.2020.9288515

Leite, J. A., Silva, D. F., Bontcheva, K., & Scarton, C. (2020). Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. *ArXiv*, 2010.04543. https://doi.org/10.48550/arXiv.2010.04543

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. *Machine Learning and Knowledge Extraction*, 285-399. https://doi.org/10.1007/978-3-030-57321-8_21

Marpaung, A., Rismala, R., & Nurrahmi, H. (2021). Hate Speech Detection in Indonesian Twitter Texts using Bidirectional Gated Recurrent Unit. *2021 13th International Conference on Knowledge and Smart Technology (KST)*, 186-190. https://doi.org/10.1109/kst51265.2021.9415760

Pamungkas, E. W., Basile, V., & Patti, V. (2020). Misogyny Detection in Twitter: A Multilingual and Cross-Domain Study. *Information Processing and Management*, *57*(6), 102360. https://doi.org/10.1016/j.ipm.2020.102360

Perifanos, K., & Goutsos, D. (2021). Multimodal Hate Speech Detection in Greek Social Media. *Multimodal Technologies and Interaction*, *5*(7), 34. https://doi.org/10.3390/mti5070034

Pratiwi, N. I., Budi, I., & Alfina, I. (2018). Hate Speech Detection on Indonesian Instagram Comments using FastText Approach. *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 447-450. https://doi.org/10.1109/icacsis.2018.8618182

Rodriguez Sanchez, F., Carrillo de Albornoz, J., & Plaza, L. (2020). Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access*, *8*, 219563-219576. https://doi.org/10.1109/access.2020.3042604

Sahu, G., Rodriguez, P., Laradji, I. H. Atighehchian, P. Vazquez, D. & Bahdanau, D. (2022). Data Augmentation for Intent Classification with Off-the-shelf Large Language Models. *Proceedings of the 4th Workshop on NLP for Conversational AI*, 47-57. https://doi.org/10.18653/v1/2022.nlp4convai-1.5

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, *8*(1), 101. https://doi.org/10.1186/s40537-021-00492-0

Sutejo, T. L., & Lestari, D. P. (2018). Indonesia Hate Speech Detection Using Deep Learning. *2018 International Conference on Asian Language Processing (IALP)*, 39-43. https://doi.org/10.1109/ialp.2018.8629154

Taradhita, D. A. N., & Putra, I. K. G. D. (2021). Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network. *Journal of ICT Research and Applications*, *14*(3), 225-239. 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. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382-6388. https://doi.org/10.18653/v1/d19-1670

Yoo, K. M., Park, D., Kang, J., Lee, S.-W., & Park, W. (2021). GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2225-2239. https://doi.org/10.18653/v1/2021.findings-emnlp.192