Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 373-382

Evaluating synonym augmentation impact on SBERT performance for Indonesian social media style classification

Jessicha Putrianingsih Pamput[1], Aindri Rizky Muthmainnah[1], Dewi Fatmarani Surianto[*1], Nur Azizah Eka Budiarti[1], Abdul Wahid[1]
Department of Informatic and Computer Engineering, Universitas Negeri Makassar, Makassar, Indonesia[1]

Article Info
Keywords:
FastText, Generation, SBERT, Social Media, Synonym Augmentation

Article history:
Received: November 11, 2025
Accepted: February 26, 2026
Published: May 01, 2026

Cite:
J. P. Pamput, A. R. Muthmainnah, D. F. Surianto, N. A. E. Budiarti, and A. Wahid, "Evaluating Synonym Augmentation Impact on SBERT Performance for Indonesian Social Media Style Classification", *KINETIK*, vol. 11, no. 2, May 2026.
https://doi.org/10.22219/kinetik.v11i2.2580

*Corresponding author.
Dewi Fatmarani Surianto
E-mail address:
dewifatmaranis@unm.ac.id

Abstract
Language on social media reflected the identity and characteristics of its users, including differences in language style between generations. Millennials and Generation Z were two dominant demographic groups in digital communication that exhibited linguistic variations, which often caused gaps in understanding during online interactions. Variations in language structure and expression posed challenges in understanding the context of cross-generational communication. Therefore, this study aimed to classify linguistic styles across generations in social media texts by combining Sentence-BERT (SBERT). FastText-based synonym augmentation in Indonesian, and Support Vector Machine (SVM) as a margin-based classification model that utilizes embedding representations from SBERT. The results showed that synonym augmentation improved model accuracy from 85% to 93%, with a similarity threshold of 0.7 providing the best balance between data diversity and semantic consistency. These findings confirmed that synonym-based augmentation and SBERT semantic adaptation were effective in capturing generational linguistic differences in informal Indonesian. This approach had the potential to be applied in other NLP tasks that required contextual understanding of social language variation, such as sentiment analysis and cross-generational dialect detection.

1. Introduction

The development of digital technology has changed the way humans communicate, including in the use of language [1]. Social media, blogs, and other digital platforms enable cross-generational interaction, information sharing, and the expression of opinions in text form [2],[3]. However, differences in cultural backgrounds and technological eras have created variations in the language structures used by different generations [4]. These changes reflect the dynamics of language evolution, which are influenced by digital trends, communication preferences, and information consumption habits in each generation. These differences in communication style often lead to gaps in understanding, especially when texts are interpreted differently by generations with different linguistic and technological backgrounds [5]. Therefore, analysing linguistic differences between generations is important for understanding potential barriers to intergenerational communication in the digital age.

Each generation has different communication characteristics, influenced by the environment and the dominance of technology at the time. Millennials or Gen Y, born between 1981 and 1996, grew up amid the transition from print to digital media, so their language style tends to be adaptive, combining formal elements with more relaxed and interactive communication [6],[7]. On the other hand, Generation Z or Gen Z, born between 1997 and 2012, are accustomed to instant communication via the internet, which shapes their language style to be more concise, expressive, and often uses abbreviations, emojis, and constantly evolving slang [8],[9]. These differences often create a gap in understanding between generations. The difference in language style between Millennials and Generation Z is increasingly apparent in digital communication [7].

Generation Z tends to use concise, expressive language influenced by social media trends, while Millennials are more adaptable, combining formal and casual language. These differences can lead to misunderstandings in various contexts, such as academic environments, the workplace, and digital marketing strategies. In addition to impacting communication effectiveness, these variations in language style also influence the dynamics of collaboration and the way information is conveyed between generations [10],[11]. These differences pose challenges in implementing cross-generational communication systems in multigenerational environments [5]. Therefore, understanding the communication patterns of both generations is crucial in creating more inclusive and effective interactions in the digital age. Thus, the development of natural language-based intelligent systems such as chatbots, virtual assistants, and recommendation systems can provide relevant and contextual responses.

Several studies have examined linguistic differences across groups and dialects. Sardila et al. (2024) identified distinctions between classical and modern Riau Malay in vocabulary, grammar, and social usage [12]. while Fitrie et al. (2023) and Olivia et al. (2022) analysed morphological and phonemic variations in regional dialects, contributing to descriptive linguistic studies [13], [14]. These studies aim to provide an objective and accurate description of the linguistic differences between dialects. Algorithmic comparisons have also been conducted, such as Suryono et al. (2025), who evaluated Levenshtein Distance and Jaro-Winkler with accuracy ranging from 42%–46% [15]. Meanwhile, Khiong (2021) and Erwina (2021) further explored structural and semantic differences between languages and dialects, enriching contrastive linguistic analysis [16], [17].

In NLP-based classification, transformer models have shown strong performance. El-Alami et al. (2022) demonstrated that fine-tuned AraBERT outperformed multilingual models in Arabic text classification [18]. while Fouadi et al. (2024) confirmed the effectiveness of dialect-specific BERT models [19]. Silva et al. (2022) reported that DistilBERT is more efficient than BERT while retaining 96% of its language capability [20]. Bello et al. (2023) improved sentiment analysis performance by combining BERT with CNN, RNN, and BiLSTM [21], and Saleh et al. (2025) achieved up to 89.6% accuracy using a transformer-based ensemble with logistic regression stacking [22]. Another study by Prottasha et al. (2022) compared the performance of the BERT model and traditional embedding in sentiment analysis in Bangla, with an accuracy of 68% [23]. while Ariyus et al. (2024) achieved 93% accuracy using Bi-LSTM and FastText [24]. Moreover, combining transformer embeddings with SVM continues to yield competitive results, such as, research by Thamrin et al. (2024) combining SBERT and SVM achieved an accuracy of 73%, higher than BERT and SVM, confirming the effectiveness of SVM in utilizing semantic embeddings [25]. Additionally, research by Roman et al. (2021) shows that linear-SVM achieves the highest accuracy of 86% in classes with sufficient training data, confirming the effectiveness of SVM in text classification [26].

Although Natural Language Processing (NLP) approaches have shown strong performance in various classification tasks, social media texts remain challenging due to informality and lexical variability [27]. Therefore, data augmentation techniques are commonly employed to enrich training data while preserving semantic meaning, enabling models to learn more diverse and adaptive linguistic representations [28]. Research by Abonizio et al. (2022) demonstrated that text augmentation methods such as EDA and back-translation improve sentiment analysis performance on imbalanced data when combined with LSTM, BERT, and SVM [29]. Meanwhile, research by Dal et al. (2025) proposed a ChatGPT-based augmentation approach that enhances classification accuracy in few-shot NLP tasks by increasing training data diversity [30]. However, most prior studies focus on general sentiment analysis or traditional linguistic aspects, while research specifically examining language style differences between Millennials and Generation Z remains limited [31].

Despite these advances, most studies focus on traditional linguistic variation or general sentiment classification. Research explicitly comparing the language styles of Millennials and Generation Z in informal Indonesian social media remains limited, particularly studies integrating transformer-based sentence embeddings with synonym-based augmentation. Furthermore, the impact of synonym-based augmentation on SBERT representations for capturing intergenerational linguistic patterns has not been widely explored [32]. SBERT is selected due to its ability to generate semantically meaningful sentence-level embeddings suitable for capturing subtle stylistic variations across generations. Therefore, this study addresses the identified gap by analysing linguistic differences between Millennials and Generation Z in informal Indonesian social media using an SBERT-based approach enhanced with synonym-based data augmentation and classified with SVM. This framework integrates semantic sentence embeddings, augmentation-driven linguistic diversity, and margin-based classification to more effectively capture intergenerational language patterns. Understanding these differences is increasingly important in academic, professional, and social media contexts [33],[34] as limited awareness of generational communication styles may hinder effective cross-generational collaboration in the digital era [27], [35], [36]. The findings are expected to contribute to a deeper understanding of intergenerational language dynamics and support the development of more adaptive and inclusive digital communication strategies.

2. Research Method

This stage explains the steps taken in the research related to the analysis of differences in language style between Millennials and Generation Z using the proposed method. The research process includes several main stages, namely data collection, data annotation, data division, pre-processing, implementation of the SBERT fine-tuning model, application of the classification model, and evaluation of results. A complete overview of the research stages is presented in Figure 1, which shows the proposed research method stages.

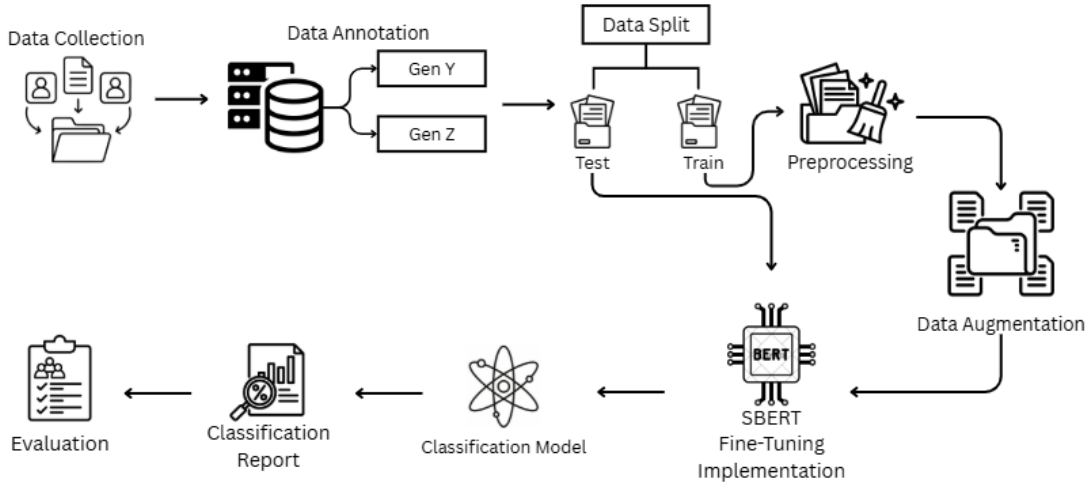

Figure 1. Stages of The Proposed Research Method

2.1 Data Collection

Data was collected manually from various social media platforms with a total of 1,850 samples, consisting of user comments on Instagram, TikTok, X, and YouTube taken from popular posts. This collection process aimed to build a dataset that represents differences in language style in the social context in Indonesia. The dataset obtained was then used to train the model to identify and classify linguistic variations more accurately. The collected data is represented in Table 1 below.

Table 1. Data Collection Results

No	Data	Class
1	*Cukup benar, bekerja dengan Gen-Z rasanya kayak naik turun... jujur saja memang mereka adalah pribadi-pribadi yang kreatif dan inovatif dalam segi pemikiran, karena banyaknya info dan trend yang mudah di akses sehingga mereka tumbuh dengan banyak ide*	Gen Y
2	*tapi untuk segi action entah kenapa mereka sangat kurang. saya rasa karena banyaknya informasi sehingga menjadikan mereka bingung untuk melangkah - takut untuk ambil keputusan. Sebenarnya sangat melengkapi generasi sebelumnya yang dimana cenderung lebih pemikir dan matang dalam setiap langkah.*	Gen Y
3	*Saya generasi Millenial, memiliki beberapa anak buah gen-z.... yang saya tangkap sejauh ini, mereka butuh pemimpin yang tegas bukan pemarah, pemberi arah dan solusi yang baik, serta bisa menjadi teman yang baik*	Gen Y
…	…	…
1848	*KIRAIIIN GUE DOANG TERNYATA BANYAK TEMENNYA* 😭	Gen Z
1849	*malaikat yang liat: buset akrab banget ni di liat liat ye*	Gen Z
1850	*nyante bgt kadang sampe rebahan malah ketiduran* 😭	Gen Z

Table 1 presents some comment data that has been categorised based on generational groups. These examples illustrate the differences in language style between Millennials and Generation Z, which serve as a reference in the annotation and linguistic analysis processes in this study.

2.2 Data Annotation

Dataset annotation was carried out to determine the category of each data point, which in this study was divided into two classes, namely Millennial Generation comments and Generation Z comments. The annotation process was carried out by analysing the linguistic characteristics of each comment to ensure accurate labelling [37]. Classification of labels is based on demographic indicators and linguistic characteristics. Comments will be classified as Millennial Generation if they have language patterns such as relatively structured sentence forms, moderate use of slang, and explanatory expressions. Meanwhile, the Generation Z label is given to comments that show characteristics of intensive

use of slang, expressive spelling variations, abbreviations, emotional markers, and informal language styles commonly observed in Indonesian social media interactions. Annotation was performed independently by two annotators, who assigned labels by evaluating language usage patterns, contextual cues derived from the respective user accounts, and the characteristics of the source content used for dataset collection. To ensure the consistency and reliability of the data annotation process, an evaluation of the level of agreement between annotators was conducted using Cohen's Kappa, which was calculated using the cohen_kappa_score function from the sklearn.metrics library. This evaluation resulted in a value of 0.9960, indicating a very high level of agreement. Mathematically, Cohen's Kappa is calculated using the following Equation 1.

$$K = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

P_o is the proportion of observed agreement, and P_e is the proportion of agreement expected to occur randomly. A kappa value close to 1 reflects that the annotation was performed consistently, making it suitable as a basis for training linguistic-based classification models.

2.3 Data Split

The dataset in this study was divided into two subsets, namely 80% training data and 20% test data [38]. This proportion was determined based on experimental results showing that this configuration provided an optimal balance between model training requirements and performance evaluation. The training data is used to build and train the model in recognising linguistic patterns between generations, while the test data serves to measure the final performance of the model in objectively classifying the language styles of Millennials and Generation Z.

2.4 Preprocessing

Preprocessing is a crucial initial stage in the classification process, as the quality of the data used greatly affects model performance [39]. In this study, preprocessing focused on cleaning and reorganising the structure of the comment text to make it more suitable for processing by the NLP algorithm. The preprocessing steps applied include case folding to convert all letters to lowercase to maintain data consistency, punctuation removal using regular expressions (regex) to eliminate irrelevant characters, and tokenisation to break the text into smaller word units. This process aims to reduce noise in the data and allow the model to focus on linguistic information that is relevant in distinguishing the language styles of Millennials and Generation Z.

2.5 Data Augmentation

The data augmentation stage aims to enrich the dataset and increase linguistic diversity in the training corpus [40]. To ensure relevant and non-extreme word representation, this study began with tokenisation and word selection based on word frequency ranging from 5 to 100 occurrences. A pre-trained FastText Indonesian language model was used to form a thesaurus based on semantic similarity between words. This dictionary is used as the basis for the augmentation process by replacing at least three tokens in each text with synonyms that have a similarity value above the similarity threshold, which is tested in the range of 0.6 to 0.8. The augmented data is combined with the original data to create a more varied combined corpus without changing the main semantic meaning of the source text.

2.6 SBERT Fine-Tuning Implementation

In this study, the SBERT (Sentence-BERT) model was applied to obtain a more in-depth semantic representation of comment texts that reflect the language style of Millennials and Generation Z. SBERT uses a transformer architecture to convert sentences or comments into fixed-dimensional vectors that include semantic information [41]. The pre-trained paraphrase-multilingual-MiniLM-L12-v2 model was used as a basis, then fine-tuned with 1,500 augmented comment data to adapt the model to the linguistic characteristics of both generations.

At this stage, the SentenceTransformer model is initialised with two main components, namely a multilingual transformer to extract text features and a pooling layer using the mean pooling method to generate sentence representations. The training data is formed in the InputExample structure (texts=[text, text], label=label) with a batch size of 16, and the loss function used is SoftmaxLoss for two classes with num_label=2. The fine-tuning process is run for 10 epochs with warmup_steps = 10 to stabilise the initial learning phase, allowing the model to recognise the unique linguistic variations of both generations in a more contextual and relevant manner.

2.7 Classification Model

Support Vector Machine (SVM) was used in this study as a classification model to distinguish the language styles between Millennials and Generation Z. SVM works by finding the optimal hyperplane that separates the two classes as much as possible [42]. In this study, the SVM classification model was implemented using the scikit-learn library through

SVC. The model was configured with a linear kernel, and class weights were adjusted using class_weight="balanced" to address class distribution imbalances in the data, enabling the model to give proportional attention to each language style across generations, including Generation Z, which tends to be more varied and complex. Details of the scenarios explored in the classification model formation can be seen in Table 2 below.

Table 2. Classification Feature Formation Scenarios

No	Scenario
1	S1 – Baseline
2	S2 – Aug
3	S3 – Aug (0.6)
4	S4 – Aug (0.7)
5	S5 – Aug (0.8)

In Table 2 above, there are five scenarios conducted to represent the linguistic characteristics of Millennials and Generation Z. The features generated from each scenario are used to transform the training and testing data, as well as form the basis for the SVM classification model training and evaluation process. The first scenario (S1) is a baseline in the form of an SBERT fine-tuning model without the application of augmentation techniques, which serves as a basic reference for assessing performance improvements in the following scenarios. The second scenario (S2) uses an SBERT fine-tuning model with synonym augmentation data. The third scenario (S3) applies SBERT fine-tuning with synonym augmentation using a minimum similarity threshold of 0.6. The fourth scenario (S4) uses a similarity threshold of 0.7, and the fifth scenario (S5) applies a higher threshold value of 0.8.

2.8 Evaluation

Model evaluation was conducted to measure classification performance in distinguishing the language styles of Millennials and Generation Z. The evaluation methods used included accuracy, precision, recall, and F1-score to assess the extent to which the model was able to accurately identify linguistic patterns of both generations. In addition, a confusion matrix was applied to analyse the distribution of predictions and identify potential biases or classification errors [43]. Furthermore, runtime measurements were evaluated to represent the efficiency of model execution time, in the form of the duration of the encoding process with SBERT, training using the SVM algorithm, and prediction on test data. The results of this evaluation form the basis for assessing the effectiveness of the model and determining further optimisation strategies, which are presented in Table 3 below.

Table 3. Confusion Matrix

		Predicted	
		Gen Y	Gen Z
Actual	Gen Y	True Positive (TP)	False Negative (FN)
	Gen Z	False Positive (FP)	True Negative (TN)

True positive (TP) refers to the number of correct predictions for the positive class, while True Negative (TN) is the number of correct predictions for the negative class. False Positive (FP) indicates the number of incorrect predictions for the positive class, while False Negative (FN) indicates the number of incorrect predictions for the negative class. Accuracy measures the extent to which the model classifies correctly, and is calculated using Equation 2 as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Equation 3 is a mathematical precision calculation that shows how many positive predictions are generated by the model.

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

Equation 4 is a mathematical calculation of recall that shows how many positive cases can be detected by the model.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Meanwhile, Equation 5 is the F1-score calculation that combines precision and recall metrics to provide a more comprehensive assessment of model performance.

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

The runtime evaluation shows the computational time efficiency required by the model in the training and prediction processes on the test data, which is calculated using Equation 6 as follows:

$$Runtime = t_{end} - t_{start} \quad (6)$$

Lower runtime values indicate more efficient computational performance, while higher values indicate increased model complexity or more intensive augmentation processes [44].

3. Results and Discussion

This study evaluates five scenarios in classifying the language styles of Millennials and Generation Z, namely SBERT Fine-Tuning, SBERT Fine-Tuning with synonym-based augmentation, and SBERT Fine-Tuning with synonym-based augmentation with a similarity threshold value variation between 0.6 and 0.8. Performance evaluation was conducted using four main metrics, namely accuracy, precision, recall, and F1-score, which were calculated separately for each class. The comparative results of the five scenarios are shown in Table 4 below.

Table 4. Research Scenario Classification Results

Scenario	Accuracy (%)	Precision (%)		Recall (%)		F1-Score (%)		Runtime (sec)
		Gen Y	Gen Z	Gen Y	Gen Z	Gen Y	Gen Z	
S1 – Baseline	85	82	87	88	81	85	84	108.38 sec
S2 – Aug	92	91	94	94	91	93	92	111.83 sec
S3 – Aug (0.6)	91	90	93	93	90	91	91	92.52 sec
S4 – Aug (0.7)	92	91	93	94	91	92	92	113.32 sec
S5 – Aug (0.8)	93	92	94	95	91	93	93	98.33 sec

Based on the results of five testing scenarios, a consistent improvement in model performance was obtained, representing a balance of performance between classes. In the first scenario, the SBERT Fine-Tuning method showed fairly stable initial performance with an accuracy of 85% and a computation time of 108.38 seconds, indicating that the fine-tuning process requires considerable resources. The performance in this scenario shows that the model is capable of recognising most of the linguistic patterns of the Gen Y and Gen Z classes, but the model still shows dependence on training data variations, where limitations in linguistic expression in the corpus can limit the model's optimal generalisation capacity. Based on the recall metric value produced, the model tends to be superior in classifying Gen Y data. This is closely related to the characteristics of the data, where the language style of the Millennial Generation tends to be more structured, consistent, and semi-formal [45], making it easier for the model to capture stable linguistic patterns. Meanwhile, Gen Z data has a high diversity of expressions, such as the use of slang, abbreviations, and variations of words with similar meanings. This is reflected in Table 5, which presents several forms of word variations commonly used by Gen Z. This wide variation remains a challenge for the model in generalising semantics effectively.

Table 5. Variations in Writing Styles by Gen Z

Basic Word	Variations in Spelling	Meaning/Usage
Banget	*bangettt, bgt, bangett, bangett banget*	*Untuk menekankan (*intensifier*)*
Kamu	*kamuu, km, kmu, kamuuh*	*Sapaan informal*
Sakit	*sakitt, sakittt, skit*	*Ekspresi sedih atau kecewa*
Please	*plis, pliss, plez*	*Permintaan secara santai/gaul*
Anjir	*anjirrr, anjirr, anjr*	*Ekspresi terkejut, kagum, atau kesal*
Parah	*parahh, parahhh, prah*	*Penekanan negatif atau positif (berlebihan)*

Several variations in the spelling of these words reflect the dynamic and expressive linguistic characteristics typical of Gen Z. Although these variations refer to the same meaning, text classification models tend to treat them as different entities due to differences in word form. This inconsistency becomes an obstacle in the process of learning

stable semantic representations, thereby reducing the model's ability to identify uniform patterns. These findings show that although fine-tuning can improve overall performance, classification results remain highly dependent on the stability and consistency of the language structure in the training corpus.

Therefore, the application of synonym augmentation in the second scenario significantly improved the model's performance. This was achieved by expanding the training corpus using semantic variations generated from a FastText-based synonym dictionary. The evaluation results showed an accuracy improvement of up to 92%. This improvement confirms that the model has become more adaptive to different expressions with similar meanings that are often used by both generations on social media. However, with the augmentation process increasing the amount of data, the computation time also increased to 111.83 sec.

The next experiment was conducted by adding a threshold parameter at the augmentation stage with the aim of controlling the quality of replacement words based on their level of semantic similarity. In the third scenario with a threshold of 0.6, the model only used synonyms with a minimum similarity level of 60%. With this threshold value, it is possible to have a wide variety of words, thereby enriching the context of the data without losing the main meaning of the sentence. The results obtained were 91% on the accuracy metric, slightly lower than the previous scenario. However, the computational time evaluation was more efficient at 92.52 sec. This shows that a low threshold produces a more varied dataset but has the potential to cause slight semantic noise due to the inclusion of synonyms with less appropriate contexts.

Increasing the threshold to 0.7 in the fourth scenario resulted in a more optimal balance between data diversity and semantic proximity. The variations formed remained contextually relevant, as the model became more selective in its choice of replacement words. The resulting accuracy reached 92% with a stable F1-score of 92%. However, increasing the threshold value also adds complexity to the synonym selection process, causing the computation time to increase to 113.32 sec. This shows a trade-off between augmentation quality and processing time efficiency, where the higher the threshold value set, the fewer synonyms meet the semantic similarity criteria. On the one hand, this can help maintain the consistency of sentence meaning and reduce semantic noise that could potentially lower linguistic context representation. On the other hand, overly strict selectivity can reduce semantic variation by reducing the model's opportunity to learn a wider range of expression styles. Therefore, determining the appropriate threshold value is a crucial factor in achieving a balance between semantic diversity that enriches the data and noise control that can reduce model stability.

In the final scenario, a threshold of 0.8 was used, whereby only words with a high level of similarity were retained in the augmentation process. This technique made the training data more semantically precise, enabling the model to identify specific linguistic patterns from both generations, Gen Y and Gen Z. Overall, the accuracy achieved was 93% with an F1-score of 93% for both classes. The significant improvement in the precision metric in this scenario indicates that the model is more accurate in its classification, thereby reducing the possibility of prediction errors between generations. However, the recorded execution time was lower at 98.33 sec, indicating that augmentation with a high similarity threshold can be more efficient because fewer synonyms are processed.

Compared to previous studies, the proposed approach demonstrates competitive performance. Research by Ariyus et al. (2024) achieved 93% accuracy using Bi-LSTM combined with FastText for Indonesian language sentiment analysis [24], while Thamrin et al. (2024) reported 73% accuracy using SBERT combined with SVM for SWOT classification [25]. In contrast, the proposed model achieved 93% accuracy specifically for intergenerational language style classification, which is inherently more complex due to overlapping vocabulary and contextual similarities between classes. This demonstrates that the integration of synonym-based augmentation with SBERT refinement effectively improves semantic discrimination in language style classification tasks.

Overall, the results of this experiment indicate that the application of synonym-based augmentation significantly improves the model's performance in recognising and understanding stylistic variations between generations. The addition of a threshold has proven to play an important role in maintaining a balance between data diversity and semantic accuracy. Higher threshold values (0.7–0.8) enable the model to produce more stable and precise linguistic representations, as the synonym selection process becomes more stringent towards the original meaning. Meanwhile, a lower threshold expands data variation with increased semantic diversity, but has the potential to cause semantic noise that can disrupt contextual coherence. Therefore, the application of augmentation with adaptive similarity settings is an effective strategy for maintaining a balance between data richness and semantic accuracy, while also improving generalisation capabilities in cross-generational linguistic classification. Thus, this approach is not only relevant to language style classification problems, but also has the potential to be adapted in various other NLP processing tasks that require a deep understanding of context and dynamic writing pattern diversity.

The results achieved in this study indicate that the classification of language styles between Millennials and Generation Z can achieve a high level of accuracy when semantic sentence embeddings are enriched with controlled lexical additions through augmentation techniques. Compared to conventional sentiment or topic classification tasks, modelling stylistic variations between generations is inherently more complex due to overlapping vocabularies and contextual similarities between age groups. Consistent improvements across various augmentation scenarios indicate

that increasing semantic diversity while maintaining contextual precision is crucial for improving classification robustness. Although the proposed approach demonstrates strong performance, the dataset is limited to informal Indonesian social media texts and two generational groups, which may restrict broader generalisation. Linguistic trends on digital platforms evolve rapidly, and stylistic overlap across age groups may vary across contexts. Future work may explore multi-task learning to incorporate social or emotional context, as well as multilingual transformer models to enhance cross-cultural and cross-platform language style classification.

4. Conclusion

This study investigates the effectiveness of SBERT enhanced with FastText-based synonym augmentation and SVM classification in modelling intergenerational language styles in Indonesian social media. The results of the experiment show a significant improvement in model performance, with accuracy increasing from 85% to 93% compared to the baseline model without augmentation. These findings confirm that expanding semantic variation through synonym augmentation contributes positively to the model's ability to understand informal and stylistic linguistic differences between generations. Using a similarity threshold of 0.7 provides the best balance between data diversity and semantic consistency, resulting in stable accuracy and F1-score of 92%, although it requires slightly higher computation time. Meanwhile, a threshold of 0.8 achieved the highest accuracy of 93%, but the high selectivity rate limited the semantic diversity in the training corpus. Overall, synonym augmentation with adaptive similarity threshold settings proved effective in enhancing semantic representation while maintaining contextual consistency, thereby strengthening the model's ability to discriminate subtle intergenerational stylistic variations in informal Indonesian social media text. However, this study still has limitations in recognising the language style of Generation Z, which tends to be more dynamic, contextual, and influenced by social media trends, making it difficult to represent semantically in an explicit manner. For further development, this approach can be expanded through multi-task learning that integrates social or emotional context into the training process. In addition, the integration of multilingual transformer models also has the potential to improve the performance of cross-cultural language style classification and more complex and realistic digital platforms.

References

[1] D. Ci. A. Ginting, S. G. Rezeki, A. A. Siregar, and Nurbaiti, "Analisis Pengaruh Jejaring Sosial Terhadap Interaksi Sosial di Era Digital," *PPIMAN: Pusat Publikasi Ilmu Manajemen*, vol. 2, no. 1, pp. 22–29, 2024. https://doi.org/10.59603/ppiman.v2i1.280

[2] C. Li, G. Ning, Y. Xia, K. Guo, and Q. Liu, "Does the Internet Bring People Closer Together or Further Apart? The Impact of Internet Usage on Interpersonal Communications," *Behavioral Sciences*, vol. 12, no. 11, p. 425, Oct. 2022. https://doi.org/10.3390/bs12110425

[3] Z. R. Eslami, T. Larina, and R. Pashmforoosh, "Identity, Politeness and Discursive Practices in a Changing World," *Russian Journal of Linguistics*, vol. 27, no. 1, pp. 7–38, 2023. https://doi.org/10.22363/2687-0088-34051

[4] A. Gondra, "Linguistic Variability across Four Generations of Basque Spanish Speakers," *Journal of Language Contact*, vol. 16, no. 4, pp. 429–455, 2023. https://doi.org/10.1163/19552629-01604001

[5] G. Šakytė-Statnickė, L. Budrytė-Ausiejienė, I. Luka, and V. Drozdova, "Internal and External Communication between Employees of Different Generations: Emerging Problems in Lithuanian, Latvian, and Swedish Tourism Organizations," 2023, *Center for International Scientific Research of VSO and VSPP*. https://doi.org/10.29036/jots.v14i26.427

[6] S. R. Febriani and A. W. Ritonga, "The Perception of Millennial Generation on Religious Moderation through Social Media in the Digital Era," *Millah: Journal of Religious Studies*, pp. 313–334, May 2022. https://doi.org/10.20885/millah.vol21.iss2.art1

[7] K. A. Boyle, "Millennial Career-identities: Reevaluating Social Identification and Intergenerational Relations," *J. Intergener. Relatsh.*, vol. 21, no. 1, pp. 89–109, 2023. https://doi.org/10.1080/15350770.2021.1945989

[8] M. Ridlo, Y. Satriyadi, A. H. Nasution, and N. A. Arandri, "Analisis Pengaruh Bahasa Gaul di Kalangan Mahasiswa Terhadap Bahasa Indonesia di Zaman Sekarang," *Jurnal Kewarganegaraan*, vol. 5, no. 2, pp. 561–569, Dec. 2021. https://doi.org/10.31316/jk.v5i2.1940

[9] N. Tarihoran, E. Fachriyah, Tressyalina, and I. R. Sumirat, "The Impact of Social Media on the Use of Code Mixing by Generation Z," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 7, pp. 54–69, 2022. https://doi.org/10.3991/ijim.v16i07.27659

[10] L. Taber, S. Dominguez, and S. Whittaker, "Ignore the Affordances; It's the Social Norms: How Millennials and Gen-Z Think About Where to Make a Post on Social Media," *Proc. ACM Hum. Comput. Interact.*, vol. 7, no. CSCW2, pp. 1–26, Sep. 2023. https://doi.org/10.1145/3610102

[11] M. D. K. Putri, B. M. K. Widarso, D. A. F. ROsanti, K. A. P. Alifariani, H. Maulana, and D. P. Arum, "Evolusi Kosa Kata Gaul Studi Antara Generasi Z Dan Milenial," *Jurnal Pustaka Cendekia Pendidikan*, vol. 2, no. 2, pp. 147–153, 2024. https://doi.org/10.70292/jpcp.v2i2.80

[12] V. Sardila, N. Faiza, Nuraini, and N. Ainiyah, "Analisis Perbedaan Bahasa Melayu Riau Klasik dan Bahasa Melayu Riau Modern di Kampar," *Gurindam: Jurnal Bahasa dan Sastra*, vol. 4, no. 1, pp. 18–26, 2024.

[13] Q. Fitrie, S. Tisnasari, and A. Supena, "Analisis Kontrastif Afiksasi Verba Bahasa Jawa Dialek Banten Dan Bahasa Indonesia Dalam Kanal Youtube Guyonan Pegandikan Periode 2021," *BAHTERA INDONESIA: Jurnal Penelitian Pendidikan Bahasa dan Sastra Indonesia*, vol. 8, no. 2, pp. 401–413, 2023.

[14] M. Olivia *et al.*, "Analisis Perbedaan Verba Dialek Sikka Natar dan Dialek Tana Ai Dalam Bahasa Sikka," *Journal Scientific of Mandalika (JSM)*, vol. 3, no. 10, 2022,

[15] W. D. Suryono, E. Utami, and D. Ariatmanto, "Analisa Perbandingan Stemming Dokumen Teks Berbahasa Jawa dengan Algoritma Levenshtein Distance Dan Jaro-Winkler," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 774–780, Jan. 2025. https://doi.org/10.29100/jipi.v10i1.6092

[16] Y. Khiong, "Analisis Perbandingan Pola Kalimat Bahasa Mandarin dengan Bahasa Indonesia," *PARAMASASTRA*, vol. 8, no. 2, pp. 180–186, 2021.

[17] E. Erwina, "Analisis Perbedaan Makna Dasar Kata Dalam Bahasa Indonesia dan Bahasa Malaysia," *SAWERIGADING*, vol. 27, no. 1, pp. 117–125, 2021.

[18] F. El-Alami, S. Ouatik El Alaoui, and N. En Nahnahi, "Contextual Semantic Embeddings Based in Fine-Tuned Arabert Model for Arabic Text Multi-Class Categorization," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8422–8428, Nov. 2022. https://doi.org/10.1016/j.jksuci.2021.02.005
[19] H. Fouadi, H. El Moubtahij, H. Lamtougui, and A. Yahyaouy, "BERT-Based Models for Classifying Multi-Dialect Arabic Texts," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 3, p. 3437, Sep. 2024. https://doi.org/10.11591/ijai.v13.i3.pp3437-3446
[20] R. Silva Barbon and A. T. Akabane, "Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study," *Sensors*, vol. 22, no. 21, Nov. 2022. https://doi.org/10.3390/s22218184
[21] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework to Sentiment Analysis of Tweets," *Sensors*, vol. 23, no. 1, p. 506, Jan. 2023. https://doi.org/10.3390/s23010506
[22] H. Saleh *et al.*, "Advancing Arabic Dialect Detection with Hybrid Stacked Transformer Models," *Front. Hum. Neurosci.*, vol. 19, 2025. https://doi.org/10.3389/fnhum.2025.1498297
[23] N. J. Prottasha *et al.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors*, vol. 22, no. 11, Jun. 2022. https://doi.org/10.3390/s22114157
[24] D. Ariyus, D. Manongga, and I. Sembiring, "Enhancing Sentiment Analysis of Indonesian Tourism Video Content Commentary on TikTok: A FastText and Bi-LSTM Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18020–18028, Dec. 2024. https://doi.org/10.48084/etasr.8859
[25] H. Thamrin, D. Oktafiani, I. I. Rasyid, and I. M. Fauzi, "Classification of SWOT Statements Employing BERT Pre-Trained Model Embedding," *Jurnal Sistem Informasi Bisnis*, vol. 14, no. 2, pp. 143–152, Apr. 2024. https://doi.org/10.21456/vol14iss2pp143-152
[26] M. Roman, A. Shahid, M. I. Uddin, Q. Hua, and S. Maqsood, "Exploiting Contextual Word Embedding of Authorship and Title of Articles for Discovering Citation Intent Classification," *Complexity*, vol. 2021, 2021. https://doi.org/10.1155/2021/5554874
[27] F. Yuni Dharta, X. Guilin, Y. Karliena, M. Butarbutar, and E. Diantoro, "Multigenerational Workforce Management Strategy in the Digital Era," *Journal Markcount Finance*, vol. 2, no. 2, 2024. https://doi.org/10.70177/jmf.v2i2.1285
[28] N. Madrueño, A. Fernández-Isabel, M. Cuesta, C. Lancho, G. Polo Vera, and I. Martín de Diego, "Novel utterance data augmentation for intent classification using large language models," *Neural Comput. Appl.*, vol. 37, no. 32, pp. 26711–26736, Nov. 2025. https://doi.org/10.1007/s00521-025-11642-3
[29] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 657–668, Oct. 2022. https://doi.org/10.1109/TAI.2021.3114390
[30] H. Dai *et al.*, "AugGPT: Leveraging ChatGPT for Text Data Augmentation," *IEEE Trans. Big Data*, vol. 11, no. 3, pp. 907–918, Jun. 2025. https://doi.org/10.1109/TBDATA.2025.3536934
[31] M. Rusydi, A. Akbar, M. Vebryanti, F. N. Tsani, and H. Z. Zavira, "Analisis Perbedaan Penggunaan Gaya Bahasa Antara Generasi Milenial dan Generasi Z dalam Komunikasi Online : Studi Kasus Akun X @xcintakiehlx dan @nnauraayu," *Jurnal Pendidikan Tambusai*, vol. 8, no. 2, pp. 27167–27175, 2024.
[32] A. Molenaar, D. Lukose, L. Brennan, E. L. Jenkins, and T. A. McCaffrey, "Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study," *J. Med. Internet Res.*, vol. 26, p. e47826, Mar. 2024. https://doi.org/10.2196/47826
[33] E. Chersoni, E. Santus, C. R. Huang, and A. Lenci, "Decoding word embeddings with brain-based semantic features," *Computational Linguistics*, vol. 47, no. 3, pp. 663–698, Oct. 2021. https://doi.org/10.1162/COLI_a_00412
[34] S. Yagi, A. Elnagar, and S. Fareh, "A benchmark for evaluating Arabic word embedding models," *Nat. Lang. Eng.*, vol. 29, no. 4, pp. 978–1003, Jul. 2023. https://doi.org/10.1017/S1351324922000444
[35] Alfa Santoso Budiwidjojo Putra, "Membangun Sinergi Lintas Generasi: Strategi Kolaboratif untuk Meningkatkan Kinerja Organisasi di Era Digital," *PaKMas: Jurnal Pengabdian Kepada Masyarakat*, vol. 4, no. 2, pp. 429–436, Nov. 2024. https://doi.org/10.54259/pakmas.v4i2.3129
[36] R. Choudhary, Y. A. Shaik, P. Yadav, and A. Rashid, "Generational Differences in Technology Behavior: A systematic Literature Review," *Journal of Infrastructure, Policy and Development*, vol. 8, no. 9, p. 6755, Sep. 2024. https://doi.org/10.24294/jipd.v8i9.6755
[37] J. C. Lapendy, A. Aulia, C. Resky, A. Tenriola, D. F. Surianto, and U. S. Sidin, "Optimizing Sentiment Analysis of Electric Vehicles Through Oversampling Techniques on Youtube Comments," *JANAPATI*, vol. 14, no. 1, pp. 169–182, 2025. https://doi.org/10.23887/janapati.v14i1.88205
[38] H. M. Alawadh, A. Alabrah, T. Meraj, and H. T. Rauf, "Attention-Enriched Mini-BERT Fake News Analyzer Using the Arabic Language," *Future Internet*, vol. 15, no. 2, Feb. 2023. https://doi.org/10.3390/fi15020044
[39] W. Wulandari *et al.*, "Semantic Feature Engineering with LSA-SVM for Cyberbullying Comment Classification on Instagram," *Informatica*, vol. 49, no. 15, Mar. 2025. https://doi.org/10.31449/inf.v49i15.6992
[40] A. Mardiah, S. Dillah, D. F. Surianto, N. Fadilah, and S. G. Zain, "Classification of Livin' by Mandiri Customer Satisfaction Using MLP with BM25 and TF-IDF Feature Weighting," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 10, no. 3, Jun. 2025. https://doi.org/10.22219/kinetik.v10i3.2248
[41] N. Fujishiro, Y. Otaki, and S. Kawachi, "Accuracy of the Sentence-BERT Semantic Search System for a Japanese Database of Closed Medical Malpractice Claims," *Applied Sciences (Switzerland)*, vol. 13, no. 6, Mar. 2023. https://doi.org/10.3390/app13064051
[42] O. Alharbi, "A Deep Learning Approach Combining CNN and Bi-LSTM with SVM Classifier for Arabic Sentiment Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 165–172, 2021. https://doi.org/10.14569/IJACSA.2021.0120618
[43] Y. Wiciaputra, J. Young, and A. Rusli, "Bilingual Text Classification in English and Indonesian via Transfer Learning using XLM-RoBERTa," *International Journal of Advances in Soft Computing and its Applications*, vol. 13, no. 3, pp. 73–87, Dec. 2021.
[44] N. Ahmed, A. L. C. Barczak, M. A. Rashid, and T. Susnjak, "Runtime Prediction of Big Data Jobs: Performance Comparison of Machine Learning Algorithms and Analytical Models," *J. Big Data*, vol. 9, no. 1, Dec. 2022. https://doi.org/10.1186/s40537-022-00623-1
[45] Z. Nassr, F. Benabbou, N. Sael, and T. Hamim, "Improving Sentiment Analysis Performance on Imbalanced Moroccan Dialect Datasets Using Resample and Feature Extraction Techniques," *Information (Switzerland)*, vol. 16, no. 1, Jan. 2025. https://doi.org/10.3390/info16010039