

Exploring the Impact of Back-Translation on BERT's Performance in Sentiment Analysis of Code-Mixed Language Data

Nisrina Hanifa Setiono^{*1}, Yunita Sari²

¹Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia

²Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia

e-mail: ^{*1}nisrinahanifasetiono@mail.ugm.ac.id, ²yunita.sari@ugm.ac.id

Abstrak

Analisis sentimen pada teks code-mixed merupakan tantangan dalam pemrosesan bahasa alami (NLP), khususnya untuk kombinasi Bahasa Indonesia dan Inggris yang sering ditemukan di media sosial seperti Twitter. Data yang bersifat informal serta keterbatasan model yang dilatih pada data formal menyebabkan performa analisis sentimen kurang optimal. Penelitian ini bertujuan menerapkan metode back translation guna mengatasi tantangan yang muncul akibat sifat informal dari data code-mixed Bahasa Indonesia-Inggris sehingga mengoptimalkan performa model BERT untuk meningkatkan akurasi analisis sentimen. Metode ini diterapkan pada dataset INDONGLISH yang terdiri dari 5.067 cuitan Twitter berlabel positif, negatif, atau netral. Hasil penelitian menunjukkan bahwa penerapan back translation langsung pada data tweet memberikan hasil lebih optimal karena mampu mempertahankan makna asli, sehingga meningkatkan performa model. Sebaliknya, ketika back translation diterapkan setelah translasi monolingual, akurasi model justru menurun akibat distorsi makna. Proses translasi berulang mengubah struktur atau konteks kalimat, menyebabkan ketidaksesuaian label sentimen. Hasil ini menunjukkan bahwa setiap tambahan proses translasi berisiko mengurangi akurasi analisis sentimen, terutama pada dataset code-mixed yang sensitif terhadap perubahan linguistik. Back translation dapat menjadi solusi untuk mengformalkan data dengan mempertahankan konteks asli, sehingga meningkatkan kualitas analisis sentimen pada teks code-mixed.

Kata kunci— Code-mixing, Sentiment Analysis, Back-Translation, BERT, Informal Text.

Abstract

Social media, particularly Twitter, has become a key platform for communication and opinion-sharing, where code mixing, the blending of multiple languages in a single sentence, is common. In Indonesia, Indonesian-English code mixing is widely used, especially in urban areas. However, sentiment analysis on code-mixed text poses challenges in natural language processing (NLP) due to the informal nature of the data and the limitations of models trained on formal text. This study applies back translation to address these challenges and optimize BERT-based sentiment analysis. The method is tested on the INDONGLISH dataset, consisting of 5,067 labeled tweets. Results show that applying back translation directly to raw tweets yields better performance by preserving original meaning, improving model accuracy. However, when back translation follows monolingual translation, accuracy declines due to semantic distortions. Repeated translation modifies sentence structure and sentiment labels, reducing reliability. These findings indicate that each additional translation step risks decreasing sentiment analysis accuracy, particularly for code-mixed datasets, which are highly sensitive to linguistic shifts. Back translation proves to be an effective approach for formalizing data while maintaining contextual integrity, enhancing sentiment analysis performance on code-mixed text.

Keywords— Code-mixing, Sentiment Analysis, Back-Translation, BERT, Informal Text.

1. INTRODUCTION

Social media has emerged as a primary platform for individuals to communicate and express opinions on various topics. One notable phenomenon, particularly on Twitter, is code-mixing, where multiple languages are used within a single sentence [1]. In Indonesia, code-mixing frequently occurs between Indonesian and English, with examples such as: “Aku lagi ngerjain tugas, but I need a break” or “Kemarin meeting-nya so fun, everyone was so engaged.” This phenomenon is predominantly found in urban areas, particularly South Jakarta [2].

Despite the widespread use of code-mixed language in digital communication, its complexity presents challenges in Natural Language Processing (NLP), especially in sentiment analysis. Code-mixed text often deviates from standard grammatical structures and lacks large annotated corpora, making it difficult for conventional models to process effectively. Prior studies achieved 76.07% accuracy in Indonesian-English code-mixed sentiment analysis using Transformer-based models [2]. Building on these findings, this research successfully replicated previous studies with slightly different results, achieving 75.96% accuracy. Other research on Indonesian-Sundanese code-mixed text using IndoBERT achieved an accuracy of 81%, though it was limited by minimal preprocessing and small dataset size [3]. Similarly, studies on Hindi-English code-mixing reported F1-scores of 0.62825, emphasizing the importance of better preprocessing techniques [4].

One of the main issues affecting sentiment analysis performance on code-mixed text is informality. Unlike monolingual corpora used to train models like BERT, code-mixed datasets often contain inconsistent structures, slang, and informal expressions, leading to degraded model accuracy. Preprocessing techniques such as emoji conversion, noise filtering, and translation have been explored to address this issue, with mixed results [5]. Additionally, text transformation techniques, such as back-translation, have been proposed to improve model performance by refining textual quality and expanding dataset diversity [6].

Back-translation, a technique in Text Style Transfer (TST), involves translating a sentence into another language and then translating it back into the original language to improve its structural consistency and clarity. This method has been applied successfully in various NLP tasks, enhancing model performance by reducing noise and formalizing text. Previous studies have demonstrated that GoogleTrans, a model capable of handling Indonesian-English translation, is particularly effective for this purpose [7]. In addition, MarianMT from Hugging Face, trained on 135 Indo-European languages and No Language Left Behind (NLLB) developed by Meta (Facebook) are also utilized.

Building on these insights, this research aims to improve Indonesian-English code-mixed sentiment analysis by applying back-translation for text formalization. This study will evaluate its impact on three top-performing sentiment analysis models from previous research [2]. By addressing data informality and enhancing textual consistency, we seek to improve accuracy benchmarks and contribute to the advancement of sentiment analysis models capable of handling complex code-mixed online discourse.

2. METHODS

The research methodology begins with acquiring labeled data, followed by preprocessing and applying back-translation to enhance data quality. The resulting dataset is then used for model training, validation, and finally, testing and evaluation. The picture of this research is carried out as in Figure 1.

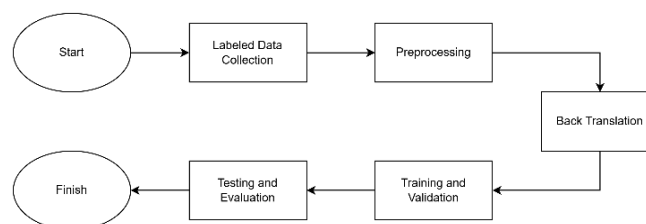


Fig. 1. Research Overview

2.1 Data Collection

This study utilizes an Indonesian-English code-mixed dataset originally constructed by Astuti et al. [2], accessible at <https://github.com/laksmitawidya/indonglish-dataset>. The dataset is associated with the sociolinguistic phenomena of Indonesian-English language usage among South Jakarta youth, as described by Wijaya and Bram [8]. Sourced from Twitter posts collected between August 2020 and September 2022, the dataset comprises 5,067 tweets. Each instance includes the original tweet, its Indonesian translation, its English translation, and sentiment labels (positive, negative, neutral) assigned by five annotators.

2.2 Preprocessing

The preprocessing stage involved multiple steps to prepare the dataset for modeling. The steps are grouped into three main processes: cleaning text, emoji conversion, and lexicon normalization. Each process is described in detail below. This follows the procedure in Astuti et al. [2] research.

2.2.1 Text Cleaning



Text Cleaning was performed to remove unnecessary or undesirable elements from the data. The steps included:

1. Profanity: Offensive words in English were removed to ensure the dataset was appropriate. Using `profanity.censor()` from `better_profanity` library.
2. Contractions: Abbreviated words in English were removed to ensure the dataset was appropriate for analysis.
3. Mentions: Any mentions (e.g., @username) were removed.
4. Hashtags: Hashtags (e.g., #example) were stripped from the text.
5. URLs: All URLs present in the text were deleted.

These steps ensured that the dataset was clean and free of extraneous information that could interfere with model training.

2.2.2 Emoji Conversion

Emojis present in the text were converted into their equivalent textual descriptions. By converting emojis into text, the sentiment and emotional context expressed in the data could be retained and interpreted effectively during analysis. For example:

-   → ":frowning_face: :frowning_face:"

2.2.3 Lexicon Normalization

Lexicon normalization aimed to standardize informal language in the dataset. This process utilized the Colloquial Words dictionary developed by Salsabila et al. [9].

2.3 Back Translation

The back-translation process begins with the collection of monolingual corpora in the target language. In this context, the monolingual corpus consists entirely of texts in a single language, such as Indonesian. This data serves as the initial input for the back-translation process. Pre-trained models from the Hugging Face library are utilized to implement this step, specifically the `Helsinki-NLP/opus-mt-id-en` model for translating texts from Indonesian to English. Once the texts are translated into English, they are translated back into Indonesian using the `Helsinki-NLP/opus-mt-en-id` model. Additionally, this study employs the GoogleTrans model and `facebook/nllb-200-distilled-600` to generate variations in the back-translated results.

The back-translated outputs are inspected to ensure no duplication with the original dataset. Furthermore, the quality of the back-translated data is evaluated using a formality index. This method ensures that the back translated data remains representative of the original dataset and meets the criteria for subsequent analytical processes. To compute the formality index,

stemming is applied using the Sastrawi library for Indonesian text and the NLTK library for English text. Both libraries are used to check whether words belong to standard Indonesian or English vocabulary. The formality index is calculated as the ratio of standard words to the total number of words in a given text. The translated samples are described in Table 1.

Table 1. Example of Translated Text Using Back Translation

Original Text	Translated Text
Saya suka coding karena it's really fun and challenging	Saya suka coding karena itu benar-benar menyenangkan dan menantang
Aku beneran gak tau apa yang dia bilang, tapi kayaknya dia really mad	Aku benar-benar tidak tahu apa yang dia katakan, tapi dia benar-benar tampaknya marah.
Memang butuh support system buat confidence lagi	Ini akan membutuhkan lebih banyak sistem pendukung untuk kepercayaan diri.

2. 4 Fine-tuning BERT

Understanding the context of both preceding and succeeding words is critical for generating robust textual representations in natural language processing (NLP). The Bidirectional Encoder Representations from Transformers, widely known as BERT, achieves this by modeling word relationships bidirectionally within a sentence. By considering all words in the surrounding text, BERT provides a deep and nuanced understanding of language, making it a powerful tool for NLP tasks [10].

Developed by Devlin et al., BERT relies on a two-phase process comprising pretraining and fine-tuning. During the pretraining phase, the model learns general language features from vast amounts of unlabeled text. This is accomplished through tasks such as predicting masked words (Masked Language Modeling) and determining the relationship between sentence pairs (Next Sentence Prediction) [10]. Fine-tuning, on the other hand, involves adapting the pretrained model to specific tasks by training it further on labeled datasets. This process enables the model to specialize in solving targeted problems with exceptional precision [10].

This research leverages the strengths of pretrained BERT models to address the challenge of sentiment analysis. A significant advantage of using pretrained models lies in bypassing the need to construct and train models from scratch, saving both time and computational resources. For this study, three key variants of BERT are employed: BERTweet, a model fine-tuned on English tweets [11]; IndoBERTweet, optimized for Indonesian tweets [12]; and Multilingual BERT, which was pretrained on a multilingual corpus from Wikipedia encompassing 104 languages. These models serve as a foundation, providing linguistic insights and contextual representations that enhance the accuracy and efficiency of sentiment analysis tasks.

2. 5 Experimental Setup

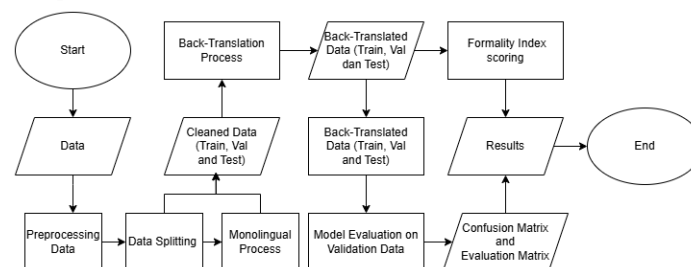


Fig. 2. Overview of Scenario 1 & 2

The flowchart above illustrates the general workflow for Scenarios 1 and 2. In Scenario 1, the process begins with data cleaning, preprocessing, and emoji conversion, followed by translation into Indonesian before applying back-translation (Indonesian-English-Indonesian). After back-translation, the formality index is computed to measure the formality of the back-

translated data. The processed data is then trained using the IndoBERTweet model, validated, and evaluated for accuracy, precision, recall, and F-measure.

In Scenario 2, the workflow is similar, but the text is first translated into English, followed by back-translation (English-Indonesian-English). Formality index is calculated, after which the data is trained using the BERTweet model, validated, and evaluated with the same metrics as Scenario 1.

Additionally, both scenarios are also tested without a monolingual setup, where no initial translation is applied, to compare performance across different preprocessing strategies.

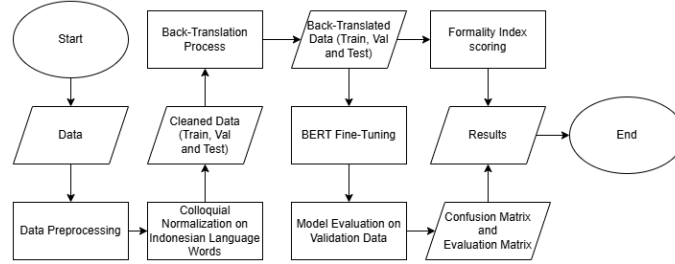


Fig. 3. Overview of Scenario 3 & 4

Fig. 3 illustrates the workflow for Scenarios 3 and 4. In Scenario 3, the process starts with data cleaning and colloquial normalization for Indonesian words. The data then undergoes back-translation (Indonesian-English-Indonesian) and is fine-tuned using the MultilingualBERT model. The model is then validated and evaluated to generate performance metrics, and the formality index is calculated to assess the formality of the back-translated data.

In Scenario 4, the steps are similar; however, the back-translation direction is English-Indonesian-English before fine-tuning with the same MultilingualBERT model. The formality index is also measured to evaluate the results of the back-translation process.

2. 6 Evaluation

2. 6.1 Index Formality

The evaluation of text formality utilized the Sastrawi library for stemming Indonesian words and the NLTK library for stemming English words. Both libraries performed checks to determine whether words in the dataset conform to standard forms of their respective languages. The formal words ratio was calculated to derive the formality index of the back-translated text.

$$\text{Index Formality} = \frac{\text{Number of Recognized Words}}{\text{Total Words}} \quad (1)$$

2. 6.1 Model Evaluation

The evaluation of this study is conducted using a Confusion Matrix to measure accuracy, precision, recall, and F1-score. Metrics such as "accuracy," "precision," and "recall" offer insights into how well the BERT model can classify the data into specific categories.

Accuracy is defined as the proportion of correctly classified samples (true positives and true negatives) to the total number of samples, as expressed in the following equation:

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}} \quad (2)$$

Precision measures the proportion of true positives to the sum of true positives and false positives, defined as

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Indicating the reliability of the model for each class. Recall, also referred to as sensitivity, evaluates the ratio of true positive predictions to the sum of true positives and false negatives, capturing the model's ability to correctly identify instances of a specific class:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Finally, the F1 Score, a harmonic mean of precision and recall, is used to balance the trade-off between these two metrics. It is calculated as:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

2. 7 Testing

Testing phase was conducted using preprocessed test data, with the back-translation process applied consistently with the procedure utilized during training. The evaluation involved comparing the predicted outputs of the model against the actual labels in the test dataset. To assess the model's performance, a confusion matrix was employed, enabling the calculation of key metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's predictive capability and overall effectiveness

3. RESULTS AND DISCUSSION

Based on the research that has been done, here are the results of the Back Translation in the first scenario. Table 2 presents the results of back-translation using the three models.

Table 2. Back Translation Scenario 1

Original Tweet	Google	MarianMT	NLLB
why people people ini kepo tingkat tinggi? Goks, ampe brightness hape dan laptop gue gelapin masih ajeeee...	mengapa orang-orang ini berada pada level yang begitu tinggi? Astaga, kecerahan hp dan laptopku pun masih gelap..	mengapa orang-orang ini tinggi tingkat kepo? Goks, ampe telepon kecerahan dan laptop saya gelap di dalam masih ajee..	Mengapa orang-orang menjaga tingkat tinggi ini?
deep talk ternyata sepenting ituuu dan selalu ngerasa terharu setelahnya:pleading_face:	Pembicaraan yang mendalam ternyata sangat penting, dan saya selalu merasa terharu setelahnya. :memohon_wajah:	Pembicaraan yang mendalam ternyata sama pentingnya dengan itu, dan selalu merasa tergerak setelah itu. :memohon_wajah:	Bicara mendalam ternyata menjadi hal yang baik , Dan selalu merasa terharu setelahnya
Monolingual Tweet	Google	MarianMT	NLLB
mengapa orang -orang ini kepo tingkat tinggi? Goks, kecerahan amp hape dan laptop gue gelapin masih ajeeee ...	kenapa orang-orang ini begitu berlevel tinggi? Ya ampun, kecerahan ponsel dan laptopku masih gelap...	Goks, ampli kecerahan dan laptopku masih gelap.	Kenapa ini tingkat tinggi?
pembicaraan mendalam ternyata memisahkan ituuu, dan selalu ngerasa terharu setelahnya. :pleading_face:	percakapan yang mendalam sebenarnya memisahkan mereka, dan selalu merasa tersentuh setelahnya. :pleading_face:	pembicaraan dalam berubah menjadi itu, dan selalu merasa tergerak setelah itu.:pleading_face:	Aku selalu merasa terharu setelahnya.

GoogleTrans adopts a more formal tone, as seen in its translation of "goks" into "yaampun." and maintaining their overall meaning. MarianMT tends to alter the original focus, such as in "mengapa orang-orang ini kepo tingkat tinggi?", where the intended criticism of people shifts to a complaint about screen brightness, changing the original meaning. Meanwhile, NLLB significantly shortens sentences, making them more concise but often at the cost of essential details. In contrast, without monolingual translation, the sentence structure remains more faithful to the original, preserving both tone and meaning with minimal distortion. The direct back translation approach ensures greater contextual accuracy, and while NLLB continues to simplify sentences, the overall meaning remains more intact compared to when monolingual translation is introduced. Table 3 presents the results of back-translation applied to scenario 2 using these models.

Table 3. Back Translation Scenario 2

Original Tweet	Google	MarianMT	NLLB
nama : nn.d hobi : ngajak overthinking bareng	Name: Ms.D Hobby: Inviting overthinking together	name: n.d hobby: invite overthinking together	Mengapa orang-orang menjaga tingkat tinggi ini?
contohnya langsung ngomong to the point "besok aku ke rumah ya buat lamar kamu"	for example, say straight to the point "tomorrow I will come to your house tomorrow to propose to you"	For example, I'll go to the point of "I'm coming home tomorrow"	For example, I'm going to talk to you right away.
Tweet Monolingual	Google	MarianMT	NLLB
name: nn.d hobby: invite overthinking together	name: nn.d hobby: inviting people to think together	name: n.d hobbies: invite overthinks together	I am not interested in this.
for example directly talking to the point "tomorrow I go to home to apply for you"	for example, get straight to the point "tomorrow I will go home to propose to you"	For example speaking directly to the point "tomorrow I go to the house to apply for you"	For example, speaking directly with the point "Tomorrow I will go home to sign up for you"

With monolingual translation, Google's back translation preserves the overall context but tends to neutralize negative sentiment. For instance, "name: n.d hobby: inviting people to think together" replaces "overthinking" with "inviting people to think," softening its original nuance. MarianMT exhibits a shift in focus, as seen in "for example, speaking directly to the point," which alters the intended meaning. Meanwhile, NLLB significantly shortens sentences, often resulting in the loss of crucial details. In contrast, without monolingual translation, Google retains "overthinking," preserving its original negative connotation. MarianMT maintains key terms while keeping the sentence structure closer to the source text. These findings suggest that direct back translation better preserves the original meaning, whereas monolingual translation introduces subtle shifts in tone and focus, potentially altering the intended message. Table 4 presents the results of back-translation applied to scenario 3 using these models.

Table 4. Back Translation Scenario 3

Original Tweet	Google	MarianMT	NLLB
literally almost everytime 😊	secara harfiah hampir setiap waktu 😊	Benar-benar hampir setiap kali	Hampir setiap kali
contohnya langsung ngomong to the point "besok aku ke rumah ya buat lamar kamu"	for example, say straight to the point "tomorrow I will come to your house tomorrow to propose to you"	For example, I'll go to the point of "I'm coming home tomorrow"	For example, I'm going to talk to you right away.

GoogleTrans maintains the closest structural alignment with the original text but may sound overly rigid. MarianMT enhances emphasis but introduces formatting inconsistencies,

affecting readability. Meanwhile, NLLB significantly shortens the translation and alters the meaning, leading to a loss of nuance. Table V presents the results of back-translation applied to scenario 4 using these models.

Table 5. Back Translation Scenario 4

Original Tweet	Google	MarianMT	NLLB
seperitnya notif whatsapp Cuma rame karena group bestie dan group kelas sama doang ya bund 😊	It's like the WhatsApp notifications are just busy because the bestie group and the class group are the same, bro 😊	Not if what happens to the group is just Rame because of bestie and the same class group doong ya bund	I'm not sure if I'm going to be able to do it.
keren sih para sutradara, produse dan para pemeran utama tarian lengger maut share lagi kak, kepo aku	It's really cool that the director, producer and main actors danced the death lengger. Share again, bro, I'm curious	“fontcolor=” # FFFF00”cool fontcolor=” # FFFF00”The directors,produce fontcolor=” # FFFF00”I leadmen fontcolor=” # FFFF00”deathlenggerdanceshareaga in fontcolor=” # FFFF00”bro, fontcolor=” # FFFF00”me”	I'm not sure if I'm going to be able to do it.

Google Translate maintains the original context more accurately than the other models. MarianMT introduces anomalies, such as repetitive phrases that distort the sentence structure and shift neutral expressions toward a more negative tone. Meanwhile, NLLB struggles with certain sentence structures, often generating generic outputs like "I'm not sure if I'm going to be able to do it," which diminishes contextual accuracy. These findings highlight the limitations of certain translation models in handling complex linguistic structures, leading to potential loss of meaning and coherence. Table VI presents the results of formality index.

Table 6. Formality Index

Raw Data Score: 0.66			
Scenario	Translation Model	Score 1	Score 2
Scenario 1	Astuti et al (2021)	0.72	
	Google	0.84	0.83
	MarianMT	0.82	0.81
	NLLB	0.88	0.87
Scenario 2	Astuti et al (2021)	0.76	
	Google	0.80	0.79
	MarianMT	0.79	0.73
	NLLB	0.85	0.83
Scenario 3	Astuti et al (2021)	0.75	
	Google	-	0.84
	MarianMT	-	0.86
	NLLB	-	0.88
Scenario 4	Google	-	0.81
	MarianMT	-	0.76
	NLLB	-	0.85

The application of back translation enhances the formality of the text compared to both the original data and monolingual translation. The original data had a formality score of only 0.66, indicating that 66% of the words were considered formal, highlighting the presence of informal elements such as slang, abbreviations, and code-mixing commonly found in raw data. Score 1 represents back translation with monolingual translation, while Score 2 represents back translation alone, with only a slight difference between them. Overall, back translation significantly improves the level of text formality across all scenarios. This improvement is consistent with findings from Astuti et al. (2021), which also showed an increase in formality scores after applying translation techniques.

Table 7. Final Results

Scenario	Translation	Translasi Model	Precision	Recall	F1-Score	Accuracy
Scenario 1 IndoBERTweet	Monolingual Astuti dkk. (2021)	Google	0.7591	0.7596	0.7572	0.7596
	<i>Back translation</i>	Google	0.7543	0.7546	0.7544	0.7546
		MarianMt	0.6769	0.6758	0.6756	0.6844
		NLLB	0.6702	0.6709	0.6685	0.6709
	Monolingual + <i>Back translation</i>	Google	0.7329	0.7349	0.7330	0.7349
		MarianMt	0.6817	0.6795	0.6790	0.6795
		NLLB	0.6090	0.6092	0.6091	0.6092
Scenario 2 BERTweet	Monolingual Astuti dkk. (2021)	Google	0.7255	0.7270	0.7255	0.7270
	<i>Back translation</i>	Google	0.7410	0.7428	0.7413	0.7457
		MarianMt	0.6460	0.6498	0.6426	0.6498
		NLLB	0.6134	0.6162	0.6126	0.6067
	Monolingual + <i>Back translation</i>	Google	0.7127	0.7161	0.7125	0.7161
		MarianMt	0.6918	0.6953	0.6915	0.6953
		NLLB	0.6208	0.6241	0.6202	0.6241
Scenario 3 Multilingual BERT	Colloquial Astuti dkk. (2021)	-	0.6677	0.6622	0.6631	0.6676
	Colloquial + <i>Back translation</i>	Google	0.6498	0.6478	0.6486	0.6478
		MarianMt	0.5482	0.5301	0.5220	0.5301
		NLLB	0.5708	0.5598	0.5476	0.5598
Scenario 4 Multilingual BERT	Colloquial + <i>Back translation</i>	Google	0.7078	0.7091	0.7080	0.7092
		MarianMt	0.6254	0.5964	0.5982	0.5964
		NLLB	0.5819	0.5628	0.5653	0.5628

The experimental findings indicate that in Scenario 1, IndoBERTweet was used as the pretrained model, with back translation and a combination of monolingual translation and back translation applied using various translation models. The results demonstrate that back translation alone consistently outperformed the monolingual-back translation combination. Among the translation models, Google Translate achieved the highest performance, yielding results comparable to previous studies, with a precision of 0.7543, recall of 0.7546, F1-score of 0.7544, and accuracy of 0.7546. In contrast, MarianMT and NLLB exhibited lower performance, with MarianMT achieving a precision of 0.6769 and NLLB 0.6702. Notably, when monolingual translation was introduced before back translation, Google Translate's accuracy slightly declined to 0.7349, compared to 0.7546 with back translation alone.

In Scenario 2, which employed BERTweet as the pretrained model, back translation using Google Translate again produced the best results, surpassing previous studies with an accuracy of 0.7457. While monolingual preprocessing before back translation improved the performance of MarianMT and NLLB, it negatively impacted Google Translate's performance. For instance, MarianMT's accuracy increased from 0.6498 (back translation only) to 0.6953 (monolingual + back translation), whereas Google Translate's accuracy decreased from 0.7457 to 0.7161. These findings suggest that while monolingual preprocessing benefits lower-performing translation models, it does not enhance models that are already highly optimized, such as Google Translate.

In Scenario 3, MultilingualBERT was evaluated using an Indonesian-English-Indonesian back translation approach, yet the results revealed a performance decline compared to the colloquial preprocessing method. Conversely, in Scenario 4, applying English → Indonesian → English back translation significantly improved MultilingualBERT's accuracy from 0.6676 (Astuti et al., 2021) to 0.7092, underscoring the importance of eliminating code-mixed elements

in multilingual text processing. Unlike prior normalization techniques that focused solely on colloquial terms, back translation standardizes mixed-language text into English, facilitating better model comprehension. Given that MultilingualBERT's pretraining corpus comprises 21% English and less than 2% Indonesian, converting code-mixed text into English enables the model to leverage its stronger linguistic representations, thereby improving processing efficiency. Additionally, back translation normalizes sentence structures and removes informal elements, such as abbreviations and slang, making the text more consistent with the model's pretraining data. In contrast, Scenario 3, which retained Indonesian as the final output, exhibited lower performance due to MultilingualBERT's limited exposure to Indonesian during pretraining. Table VIII presents the sentiment prediction results using the Google model without monolingual translation, compared to the findings from Scenario 2 in the study conducted by Astuti et al. (2021).

Table 8. Difference in Predicted Label

Original Tweet	Scenario Results	Label	Predicted Label
Bestie premium maksudnya gmn y kak	Scenario 2 Astuti dkk (2021) Bestie premium maksudnya gmn y sis	Neutral	Negative
	Scenario 2 BT Google What about bestie premium?	Neutral	Neutral
When most people say that time is money, I have to disagree because it isn't... Time is priceless... When you lose some money, it'll get back to you eventually... But when you lose even a second of your time, you'll never get it back...,"	Scenario 2 Astuti dkk (2021) when most people say that time is money, i have to disagree because it is ... time is priceless ... when you lose some money, thats getting back to you eventually ... but when you lose even a second of your time, youll never get it back ...	Negative	Neutral
	Scenario 2 BT Google when most people say that time is money, i have to disagree because it isn't... time is precious... when you lose some money, eventually it will come back to you... but when you lose even a little bit of your time , you will never get it back..	Negative	Negative

The following example demonstrates that back translation using Google Translate, without additional monolingual translation, results in label predictions that align with the original labels. In the first case, where the original tweet is neutral, back translation with Google incorrectly shifts the predicted label to negative. However, in the second case, where the original label is negative, back translation with Google successfully preserves the intended sentiment, ensuring the predicted label remains negative. This suggests that while back translation alone can effectively maintain the sentiment in certain cases, additional monolingual translation may introduce shifts in meaning that affect label prediction accuracy.

Overall, these findings demonstrate that back translation plays a crucial role in enhancing model performance, particularly when utilizing Google Translate. While monolingual preprocessing enhances results for lower-performing translation models such as MarianMT and NLLB, it does not yield additional benefits for high-quality translation models like Google Translate. These results suggest that the effectiveness of back translation varies based on the translation model and preprocessing approach, with Google Translate consistently delivering the most optimal performance.

4. CONCLUSIONS

Based on the findings of this study, it can be concluded that the application of the back translation (BT) method using Google on code-mixed data yields optimal results when applied

directly to raw data without prior monolingual translation. This is because Google's BT effectively preserves the original context and meaning of sentences, leading to a significant improvement in sentiment analysis model performance. However, when BT is conducted after monolingual translation, the model's performance declines due to meaning distortions. Repeated translation processes, from code-mixed to monolingual and then through BT, can inadvertently modify sentence structure or context, resulting in shifts in interpretation. Consequently, initial labeling of the text often becomes inaccurate, as a sentence initially carrying a positive sentiment may be misinterpreted as negative after multiple translation steps. These findings highlight that each additional translation process poses a risk of reducing data accuracy, particularly in code-mixed datasets, which are highly sensitive to linguistic context shifts.

REFERENCES

- [1] Patwardhan, V., Takawane, G., Kelkar, N., Gaikwad, O., Saraf, R., & Sonawane, S. (2023). Analysing The Sentiments Of Marathi-English Code-Mixed Social Media Data Using Machine Learning Techniques. 2023 International Conference on Emerging Smart Computing and Informatics, ESCI 2023. <https://doi.org/10.1109/ESCI56872.2023.10100304>
- [2] Widya Astuti, L., & Sari, Y. (2023). Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data. In IJACSA) International Journal of Advanced Computer Science and Applications (Vol. 14, Issue 10). www.ijacsa.thesai.org
- [3] Najiha, H., & Romadhony, A. (2023). Sentiment Analysis on Indonesian-Sundanese Code-Mixed Data. 2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023. <https://doi.org/10.1109/I2CT57861.2023.10126254>
- [4] Patil, A., Patwardhan, V., Phaltankar, A., Takawane, G., & Joshi, R. (2023). Comparative Study of Pre-Trained BERT Models for Code-Mixed Hindi-English Data. 2023 IEEE 8th International Conference for Convergence in Technology, I2CT 2023. <https://doi.org/10.1109/I2CT57861.2023.10126273>.
- [5] Pota, M., Ventura, M., Catelli, R., & Esposito, M. (2021). An effective bert-based pipeline for twitter sentiment analysis: A case study in Italian. *Sensors (Switzerland)*, 21(1), 1–21. <https://doi.org/10.3390/s21010133>.
- [6] Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00492-0>.
- [7] Sari, Y., & Al Faridzi, F. P. (2023). Unsupervised Text Style Transfer for Authorship Obfuscation in Bahasa Indonesia. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 17(1), 23. <https://doi.org/10.22146/ijccs.79623>.
- [8] Diva Wijaya, A., & Bram, B. (2021). A SOCIOLOGICAL ANALYSIS OF INDOGLISH PHENOMENON IN SOUTH JAKARTA (Vol. 4, Issue 4). www.news.okezone.com
- [9] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 236–239, doi: 10.1109/IALP.2018.8629151.
- [10] Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (n.d.). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://github.com/tensorflow/tensor2tensor>
- [11] N. L. Pham and V. V. Nguyen, "Adapting Neural Machine Translation for English-Vietnamese using Google Translate system for Back-translation," 2019 International Conference on Advanced Computing and Applications (ACOMP), 2019, pp. 1-6.
- [12] Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. <http://arxiv.org/abs/2011.00677>