VOL. 23, NO. 1, 2024, 96-102
elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

Language Threshold for Multilingual Sentiment Analysis System

Yuheng Kit[1] and **Musa Mohd Mokji**[1]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai Johor, Malaysia

[*]Corresponding author: yhkit2@graduate.utm.my

Abstract: Code-mixed sentences are very common in social media platforms especially in countries such as Malaysia that have more than 2 speaking languages. Although multilingual Bidirectional Encoder Representations from Transformers (mBERT) has the capability of understanding multilingualism, the sentence embeddings obtained from mBERT can be very complex for a code-mixed sentence. This is a challenge in Natural Language processing when processing informal social media text due to its complexity, especially in mixed languages like Malay-English where there is an insufficient amount of training datasets available. Thus, this paper proposes a language threshold to translate the affected words or sentence into a single language sentence and relabel the language of the sentence. The result shows an increase of 8% in accuracy when translating affected words in a sentence at the 60% language threshold using SEC PCA-200.

Keywords: Natural language processing, Sentiment analysis

Article History: received 15 April 2023; accepted 26 September 2023; published 28 April 2024.

1. INTRODUCTION

In recent years, the Malaysian government has been paying close attention to the potential of artificial intelligence (AI) and has launched the "Artificial Intelligence Roadmap 2021–2025 (AI-RMAP)" [1] to accelerate the growth and adoption of AI technology in Malaysia. The AI-RMAP initiative includes several proposed projects, such as Natural Language Processing (NLP), which aims to enhance communication between humans and machines.

The rising popularity of social media in Malaysia has been noted in various reports, including the "Malaysia Digital Economy Blueprint" report, which states that 81% of Malaysians were active on social media in 2020 [2]. This figure is expected to increase as social media becomes increasingly integral to daily life in Malaysia. Twitter is among the world's most popular social media platforms, with 368 million users globally [3] and 4.4 million active users in Malaysia alone [4].

Given these trends, the Malaysian government's focus on AI technology and NLP projects is a step in the right direction. As more Malaysians turn to social media for communication and engagement, the potential applications of NLP and other AI technologies are significant. The AI-RMAP initiative is expected to drive innovation and growth in the Malaysian digital economy and position Malaysia as a leading player in the global AI landscape.

Code-switching or code-mixing, also known as "bahasa rojak" in Malaysian linguistics, is a common informal language used in social media content in Malaysia. The Jacobson code-mixing model, created in 1996, includes five categories of code-mixed spoken interaction: monolingual English (E-), main language English with some Malay (ML-E), equal language alternation of Malay and English (=LA), main language Malay with some English (ML-M), and monolingual Malay (M-). Understanding code-mixing patterns is important when analyzing social media content in Malaysia, where 81% of the population is active on social media. The Jacobson model provides a framework for identifying and categorizing different types of code-mixed interactions, helping researchers gain a better understanding of language dynamics in social media [5].

Analyzing code-mixed text presents unique challenges for natural language processing (NLP) applications. One critical aspect is accurately identifying the language used in the text, which is essential for sentiment analysis and machine translation. However, this can be challenging due to the complexity of the code-mixed text. Researchers have identified language identification as one of the five key focus areas in analyzing this type of text, with a system that can be divided into four levels: document, sentence, word, and sub-word levels. Language tags at the document or sentence level alone are insufficient for accurately identifying the language in the code-mixed text, as mixed languages can cause language detector systems to fail. To overcome this challenge, researchers have developed methods that work at the word or sub-word level to improve language identification accuracy [6].

Code-mixed text presents a challenge due to the use of ambiguous words, which have different meanings in each language. While some words, such as "main," "fail," "cap," "jam," "atom," and "air," can be found in both Malay and English, there are many more words that are unique to each language. Researchers have proposed a lexicon-based approach, using tools such as SentiNetWord and VADER with pre-processing techniques to translate Indonesian and Javanese languages to English [7]. However, the success of this approach depends heavily on the language model used. Creating a sentiment lexicon or dictionary [8] is time-consuming, as it requires human annotation.

In multilingual sentiment analysis, sentence-level language identification is crucial to classify code-mixed text into a single language and reduce system complexity for NLP applications. Identifying code-mixed sentences alone is insufficient, and extending word-level language identification to the sentence-level can improve the accuracy of sentiment classification in multilingual sentiment analysis.

To address the challenge of the code-mixed text, this paper proposes a language threshold approach using the multilingual sentiment analysis system proposed in [9]. The solution involves performing text translation and language relabeling based on a threshold for the input sentence. This approach reduces the complexity of the classifiers by focusing on one language for each classifier. By using this approach, we can improve the accuracy of sentiment analysis in code-mixed text and reduce the computational cost of training multilingual sentiment analysis models.

2. METHODOLOGY

2.1 Language Threshold

Language threshold can be used to determine the language of the sentence or convert a sentence into a target language in a more effective way rather than depending on the language model solely.

2.1.1 Words / Sentence Translation

When a sentence is code-mixed with multiple languages, it can increase the complexity of the model during training. However, this complexity can be reduced by translating the affected words or sentences into a target language using a specified language threshold. By doing so, the sentence is simplified, and the classifier's complexity is reduced.

To achieve this, the sentence is first divided into individual words, and a language detector is then used to count the number of English and Malay words in the sentence. Next, the percentages of each language in the sentence are calculated. If the percentage of a language is greater than or equal to the specified threshold but not 100%, the sentence is classified as either English or Malay. Finally, the affected words or sentences are translated to the target language using a language detector, as shown in Algorithm 1. This process results in accurate classification and translation of code-mixed sentences, simplifying the training process and improving the model's accuracy.

Algorithm 1. Words / Sentence Translation with Language Threshold

```
Set Ne and Nm to 0
Set T to the desired language threshold percentage
For each word in the sentence:
    If the word is English:
        Increment Ne by 1
    Else:
        Increment Nm by 1
Compute %Ne = (Ne/(Ne+Nm)) * 100
Compute %Nm = (Nm/(Ne+Nm)) * 100
If T≤%Ne≤100%:
    Translate all affected words in the sentence / whole sentence to English
Else if T≤%Nm≤100%:
    Translate all affected words in the sentence / whole sentence to Malay
Else:
    Do nothing
```

2.1.2 Language Relabeling

When a code-mixed sentence is fed into the system [9], it is necessary to accurately classify the language of the sentence. Relying solely on the language classification provided by Twitter API may not be reliable, as it is prone to errors.

Algorithm 2. Language Relabeling with Language Threshold

```
Set Ne and Nm to 0
Set T to the desired language threshold percentage
For each word in the sentence:
    If the word is English:
        Increment Ne by 1
Else:
    Increment Nm by 1
Compute %Ne = (Ne/(Ne+Nm)) * 100
Compute %Nm = (Nm/(Ne+Nm)) * 100
If T≤%Ne≤ 100%:
    Label the language of sentence to English
Else if T≤%Nm≤100%:
    Label the language of sentence to Malay
Else:
    Do nothing
```

The percentages of English and Malay are calculated after detecting the language of words in the sentence with a language detector. When the percentage of language is greater or equal than the threshold and not 100%, it will be classified either into English or Malay. If the condition does not satisfy, it will use the original language as illustrated in Algorithm 2.

2.2 Types Of Language Threshold in Words / Sentence Translation

To translate sentences with mixed languages, a language translator can be used to translate the relevant words, while a language detector with a language threshold can be used to classify the sentence based on its language percentages. For example, if a sentence such as "I sangat happy" contains 2 English words and 1 Malay word with language percentages of 66.67% for English and 33.33% for Malay, it can be classified as English if the language threshold is set to 60%. In this case, the Malay word "sangat" can be translated to "very" in English.

To determine the length of a sentence, certain case-sensitive main subjects, such as "lazada", "shopee", etc., can be excluded. For instance, in the sentence "I buy my baju baru in Shopee", the sentence length will be 6 (excluding "Shopee"), which will affect the language percentages, making them 66.67% for English and 33.33% for Malay.

If the entire sentence needs to be translated, the language detector and translator can be used together. For instance, the sentence "I sangat happy" can be classified as English at 60% language threshold and translated to "I am very happy". This method can help translate sentences with mixed languages accurately and efficiently without using unnecessary words.

2.3 Multilingual Sentiment Analysis System

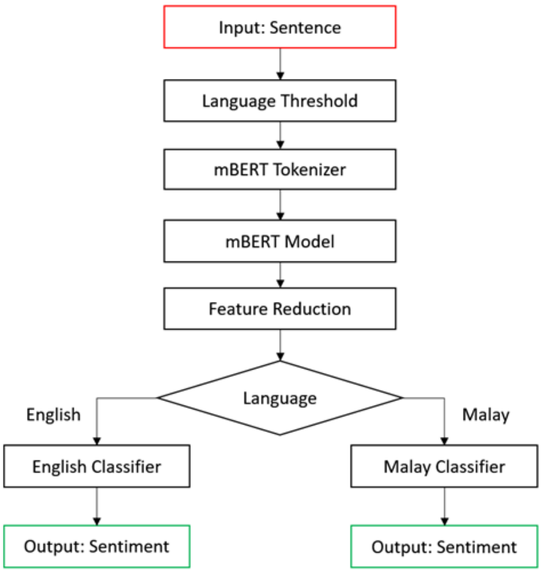

Figure 1. Multilingual Sentiment Analysis System

The multilingual sentiment analysis system proposed by [9], which utilized the Sentence Embedding Classification (SEC) strategy, demonstrated similar performance in the sentiment classification task by requiring less training time and consuming less GPU memory when compared to the fine-tuning method. Thus, this system is employed in this paper, along with an additional Language Threshold process as described in Section 2.1 and illustrated in Figure 1. The input sentence will be processed by the language threshold process to obtain the language of the sentence and translate the affected words or sentence into the target language. Next, the new input sentence will be fed into multilingual Bidirectional Encoder Representations from Transformers (mBERT) tokenizer to tokenize the sentence, Then, tokens will be fed into a mBERT [10] model to obtain the sentence embeddings from the last layer of [CLS] token [9]. Next, the sentence embeddings will be further reduced using feature reduction algorithms to get reduced features and feed them into the classifier depending on language to get the final sentiment. The sentence embedding classification with the original embeddings from the CLS token, which is 768 sentence embeddings, is called SEC SE-768. The sentence embedding classification with the sentence embeddings reduced from 768 to 200 using Principal Component Analysis (PCA) as a feature reduction algorithm is called SEC PCA-200 [9].

3. RESULT AND DISCUSSION

3.1 Data Collection

The dataset used in this study was collected from Twitter's official API, which was accessed under an academic research license. The focus of the dataset was on two popular e-commerce platforms, "Lazada" and "Shopee," during the period from 1.1.2020 to 23.10.2022 in Malaysia. Since Twitter's API does not support the Malay language, Indonesian was used as a query language.

The dataset consists of 20,000 data points each for English and Malay. The sentiment in both datasets was labelled with three classes - Negative (0), Neutral (1), and Positive (2) - with the assistance of a certified e-commerce industry expert. The distribution of classes within the dataset can be found in Table 1.

Overall, this dataset provides valuable insights into consumer sentiment towards Lazada and Shopee in Malaysia during the specified period. However, it is important to take note of the limitations of the dataset, such as the fact that the sentiment labels were assigned by only one expert and the use of Indonesian as the query language.

Only 15k data in each dataset is labelled, which is due to duplicate data, and part of it serves as the testing dataset. The English dataset contains 300 samples per class, whereas the Malay dataset has 1000 samples per class, resulting in an imbalance of data, as shown in Table 2. For the testing dataset, a total of 150 samples are included, consisting of a mix of both English and Malay data.

Table 1. Class distribution of the 15K dataset

	Negative	**Neutral**	**Positive**
English	332	14318	350
Malay	1077	12882	1041

Table 2. Class distribution of the experiment dataset

	Negative	**Neutral**	**Positive**
English	300	300	300
Malay	1000	1000	1000

The following presents a list of the various types of sentiments and the annotation rules used in the dataset. It categorizes sentiments as Positive, Neutral, or Negative, and outlines the guidelines for annotating each category:

1. **Positive Sentiment:** Sentences that represent happiness, satisfaction, celebration, delight, appreciation, interest towards the subject, admiration and any related to positive category [11].
   a. Any sentence that expresses positive meaning [11] [12] [13].
   b. Any sentence that consists of more positive meaning than negative meaning.

c. Any sentence that consists of both neutral and positive meaning [13].

For example,
English Sentence: "I bought the landscape painting at shopee 😭 so cute"
Description: The author likes the product by describing the product as cute.
Malay Sentence: "Yang ni murah lagi"
Description: The English version of this sentence is "This is cheaper". The author shares cheaper options of the product to others and carries a positive attitude towards it due to the product being cheaper.

2. **Neutral Sentiment:** Sentences that do not describe speaker's either positive or negative sentiments, then the sentence belongs to the neutral/mixed category [13].
   a. Any question or answer that does not carry any positive or negative meaning [11] [12].
   b. Any suggestion / comment / assumption / sharing that does not carry any positive or negative meaning [11] [12] [13].
   c. Any sentence that has equal positive and negative meaning or conflict [11].
   d. Any sentence that contains factual information [13].
   e. Any sentence that has a promotion element.

For example,
English Sentence: "i don't know you, but i want you 🎶purchase me, the Tinted Lip Balm @skineats_ then i'll be your best friend! ♡ just like this beautiful soul! 🛒: https://t.co/LqjO6Sa2Ni or https://t.co/n37c0DgnAh https://t.co/VmyFpe1N2Z"
Description: The author promotes a product with product link and positive terms such as 'beautiful' and 'best'. However, the sentence is neutral sentiment as it is more of a promotional sentence.
Malay Sentence: "@tattyhassan Ada jual kat shopee rasanya"
Description: The english version of this sentence is "Should be selling in shopee". The author shares information to others without carrying any sentiments.

3. **Negative Sentiment:** Sentences that express negative emotions such as anger, disgust, envy, irritation, and so on.
   a. Any sentence that has negative sentiment in all aspects without containing negation [13].
   b. Any sentence with more negative meaning than positive / neutral meaning [13].
   c. Any sentence that contains disagreement or rejection [12] [13].
   d. Any sentence that contains irony or sarcasm [12].
   e. Any sentence that has a complaint element [12].

For example,
English Sentence: "Why put name Shopee EXPRESS when the service is slow n shitty @ShopeeMY?"
Description: The author expresses the bad feelings towards Shopee Express.
Malay Sentence: "Bising la shopee ni!"
Description: The english version of this sentence is "This shopee is noisy!". The author expresses negative feelings towards Shopee.

3.2 Language Threshold

In evaluating the English classifier, the accuracies are determined through training and testing on an English dataset. Similarly, the Malay classifier's accuracies are determined using a Malay dataset. The training time for each classifier is presented in the table as well. To assess the overall system's performance, a separate testing dataset described in Section 3.1 is employed and the accuracy is denoted as "Final" as well as the accuracy of the language detector is reported by utilizing the Google Translate API. Table 3 shows the accuracy of the original dataset without any text translation and language labeling. The accuracy for Sentence Embedding Classification (SEC SE-768) is 58.67% while the accuracy for Sentence Embedding Classification with PCA 200 principal components (SEC PCA-200) is 60%. Figure 2 and Figure 3 show the heatmaps for SEC SE-768 and SEC PCA-200, respectively, using the testing dataset described in Section 3.1. Referring to heatmap in Figure 3, there are improvements in prediction with SEC PCA-200 as compared to SEC SE-768 with lesser features which is from 768 features reduced to 200 features.

Table 3. Accuracy for origins dataset

Process	SEC SE-768	SEC PCA-200
Original	English Classifier: 63.56% (0.09s) Malay Classifier: 62.53% (1.06s) **Final**: 58.67% (Language detector: 88.00%)	English Classifier: 60.89% (0.22s) Malay Classifier: 61.73% (0.83s) **Final**: 60.00% (Language detector: 88.00%)

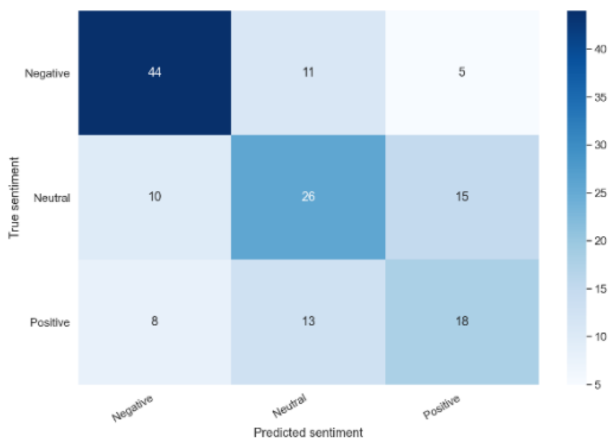

Figure 2. Heatmap for SEC SE-768 in Original Dataset

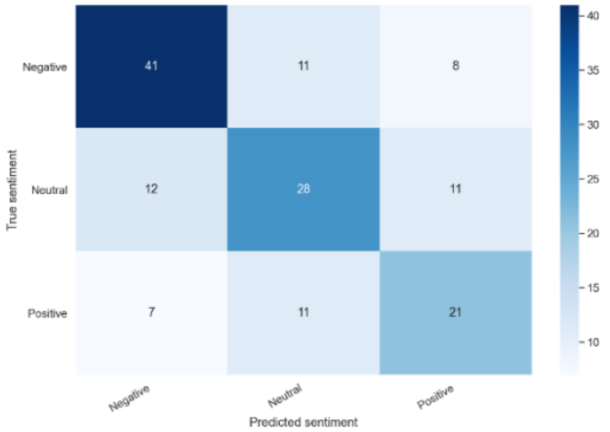

Figure 3. Heatmap for SEC PCA-200 in Original Dataset

Table 4 shows the class distribution of the training dataset after language relabeling with language threshold in word level and sentence level. The dataset sizes are different from the training dataset and the classes are unbalanced.

Table 4. Dataset for Language Relabeling with Language Threshold (Same for Word & Sentence Level)

Threshold	Dataset	Negative	Neutral	Positive
90%	English	300	302	300
	Malay	1000	998	1000
80%	English	291	310	307
	Malay	1009	990	993
70%	English	281	352	332
	Malay	1019	948	968
60%	English	281	441	436
	Malay	1019	859	864

Table 5 shows the class distribution of the training dataset after language relabeling with language threshold in word level and sentence level with main subjects described in Section 2.2. The dataset sizes are different from the training dataset and the classes are unbalanced.

Table 5. Dataset for Language Relabeling with Language Threshold (Same for Word & Sentence Level with Main Subjects)

Threshold	Dataset	Negative	Neutral	Positive
90%	English	300	302	300
	Malay	1000	998	1000
80%	English	301	314	310
	Malay	999	986	990
70%	English	301	365	345
	Malay	999	935	955
60%	English	337	465	455
	Malay	963	835	845

Table 6. Accuracies of Language Threshold using **Word Level** for Text Translation, Language Relabelling and Combination.

Threshold	Text Translation		Language Relabeling		Combination	
	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)
90%	**63.33**	62.67	57.33	64.00	57.33	64.00
80%	62.67	64.67	61.33	**64.67**	57.33	**66.00**
70%	58.67	62.67	60.00	60.67	**61.33**	60.00
60%	62.00	**68.00**	**62.00**	61.33	60.00	62.00

The results showed that the accuracy of text translation improves when a language threshold was applied. The original accuracy for SEC and SEC PCA-200 were 58.67% and 60.00%, respectively, as seen in Table 6. However, with a 60% language threshold, the accuracy for SEC PCA-200 increased by 8%, indicating the importance of translating code-mixed sentences into single language sentences. This simplified the sentence embeddings obtained from mBERT, allowing the classifier to better understand the features for classification.

Despite the discrepancies observed in the training dataset between Table 5 and Table 4, the accuracy levels for language relabeling and combination (comprising text translation and language relabeling) remained generally higher than the baseline accuracies, as demonstrated in Table 6. In Table 6, it was evident that both the accuracies of SEC SE-768 and SEC PCA-200 in language relabeling and combination exhibited improvements, particularly at the 60% language threshold. It was anticipated that enhanced accuracy would be attained with a more balanced training dataset and different configuration of classifier training. Furthermore, the importance of the language threshold in language relabeling persists, as real-world sentences may not always come with language labels.

In comparing the accuracies for a 60% language threshold in text translation experiments between Table 6 and Table 7, it was evident that word-level translation outperformed sentence-level translation. Specifically, word-level translation achieved an accuracy of 68%, while sentence-level translation achieved 61.33% when using SEC PCA-200, suggesting that it is not necessary to translate the entire sentence but only the affected words. This may be due to the fact that translating the whole sentence may result in grammar errors and inappropriate word choices, while mBERT does not consider stop words or unimportant words. Therefore, translating affected words is a better option, as it is faster than translating at the sentence level.

The study also included experiments on word-level and sentence-level analysis with main subjects, and the outcomes were presented in Table 8 and Table 9 correspondingly. The omission of the count for main subjects could have affected the language proportion in a sentence; however, since not all sentences contained those main subjects, it had an insignificant impact on accuracy.

Notably, the combination method achieved an accuracy of 70.67% for SEC PCA-200 as demonstrated in Table 8.

Table 7. Accuracies of Language Threshold using **Sentence Level** for Text Translation, Language Relabelling and Combination

Threshold	Text Translation		Language Relabeling		Combination	
	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)
90%	60.00	58.67	**60.67**	**64.67**	58.00	61.33
80%	62.00	61.33	60.00	63.33	58.00	**66.67**
70%	63.33	**62.67**	58.00	60.00	61.33	62.00
60%	**63.33**	61.33	**60.67**	64.00	**63.33**	56.67

Table 8. Accuracies of Language Threshold using **Word Level with Main Subjects** for Text Translation, Language Relabelling and Combination.

Threshold	Text Translation		Language Relabeling		Combination	
	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)
90%	**63.33**	64.00	56.00	60.67	54.67	59.33
80%	62.67	65.33	**60.00**	63.33	**60.67**	63.33
70%	60.00	64.00	58.00	**66.67**	56.00	66.00
60%	56.67	**66.67**	58.67	57.33	60.00	**70.67**

Table 9. Accuracies of Language Threshold using **Sentence Level with Main Subjects** for Text Translation, Language Relabelling and Combination.

Threshold	Text Translation		Language Relabeling		Combination	
	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)	SEC SE-768 (%)	SEC PCA-200PC (%)
90%	62.00	64.00	**60.67**	64.67	58.67	64.00
80%	**64.00**	**66.67**	55.33	64.00	58.67	**66.67**
70%	62.00	63.33	58.67	**66.00**	58.00	62.67
60%	**64.00**	**66.67**	57.33	55.33	58.67	60.67

4. CONCLUSION

This paper examines the impact of language threshold on sentence embeddings by exploring the use of translation, language relabeling, and a combination of both techniques. The aim is to investigate how these methods can affect the complexity and accuracy of mixed-language sentences.

The findings indicate that translating code-mixed sentences into a single language can lead to simpler sentence embeddings. Additionally, increasing the language threshold generally improves the accuracy of language detection. It is worth noting that translating only the affected words in a sentence requires less processing power than translating the entire sentence, particularly when using multilingual BERT (mBERT).

Future research could investigate additional datasets using the same methods to further explore the impact of language threshold on sentence embeddings. Additionally, given the prevalence of code-mixed text in social media, simplifying input sentences from social media could be a promising area of research. By reducing the length and complexity of these sentences, the efficiency and accuracy of language detection algorithms for this type of text can be improved.

REFERENCES

[1] "Artificial intelligence roadmap 2021–2025 (AI-RMAP)," Ministry of Science, Technology & Innovation (MOSTI), 2021.

[2] "Malaysia digital economy blueprint," Economic Planning Unit, Prime Minister's Department, 2021.

[3] S. Dixon, "Number of Twitter users worldwide from 2019 to 2024," Statista - *The Statistics Platform.* Retrieved March 10, 2023 from https://www.statista.com/statistics/303681/twitter-users-worldwide/.

[4] S. Dixon, "Leading countries based on number of Twitter users as of January 2022," Statista - *The Statistics Platform.* Retrieved March 10, 2023 from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[5] N. Iman et al., "English-Malay code-mixing innovation in Facebook among malaysian university students," Researchers World - Journal of Arts, Science & Commerce, 2015, doi: 10.18843/rwjasc/v6i4/01.

[6] A. F. Hidayatullah, A. Qazi, D. T. C. Lai and R. A. Apong, "A systematic review on language

identification of code-mixed text: techniques, data availability, challenges, and framework development," in IEEE Access, vol. 10, pp. 122812-122831, 2022, doi: 10.1109/ACCESS.2022.3223703.

[7] C Tho et al., "Code-mixed sentiment analysis of Indonesian language and Javanese language using lexicon based approach," Journal of Physics: Conference Series 1869 012084, 2021, doi:10.1088/1742-6596/1869/1/012084.

[8] N.H Mahadzir1 et al., "Sentiment analysis of code-mixed Text: A review," Turkish Journal of Computer and Mathematics Education Vol.12 No.3 (2021), p.p. 2469-2478, 2021, doi: 10.17762/turcomat.v12i3.1239.

[9] Y. Kit and M. M. Mokji, "Sentiment analysis using pre-Trained language model With no fine-tuning and less resource," in IEEE Access, vol. 10, pp. 107056-107065, 2022, doi: 10.1109/ACCESS.2022.3212367.

[10] J. Devlin, M. Chang, K. Lee, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, arXiv:1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805

[11] M. S. H. Mukta et al., "A comprehensive guideline for Bengali sentiment annotation," ACM Transactions on Asian and Low-Resource Language Information Processing Vol. 21 No. 2, p.p 1-19, 2021, doi: 10.1145/3474363.

[12] J. Urpay-Camasi, J. Garcia-Calderon and P. Shiguihara, "A method to construct guidelines for Spanish comments annotation for sentiment analysis," 2021 IEEE Sciences and Humanities International Research Conference (SHIRCON), Lima, Peru, pp. 1-4, 2021 doi: 10.1109/SHIRCON53068.2021.9652313.

[13] L. Khan, A. Amjad, N. Ashraf and H. Chang, "Multi-class sentiment analysis of urdu text using multilingual BERT," Scientific Report 12:5436, 2022, doi: 10.1038/s41598-022-09381-9.