Penerbit
Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

Semi-Automatic Sentiment Identification for Malay-English Code-Switched Data

Afifah Mohd Shamsuddin[1], Sarah Samson Juan[1,*], Stephanie Chua[1], Arif Bramantoro[2]

[1] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2] School of Computing and Informatics, Universiti Teknologi Brunei, BE1410, Bandar Seri Begawan, Brunei Darussalam

ARTICLE INFO

Article history:
Received 2 August 2024
Received in revised form 18 October 2024
Accepted 16 December 2024
Available online 31 December 2024

Keywords:
Sentiment analysis; code-switching; sentiment classification; natural language processing

ABSTRACT

Informal online communication on social media platforms like Twitter, Facebook, and YouTube often involves code-switching between languages, notably Malay and English, in Malaysia due to its diverse society. This phenomenon challenges sentiment analysis tasks, as the intermixing of languages within sentences or phrases increases the likelihood of inaccurately classified sentiment. Sentiment analysers built with models trained with monolingual data will cause misclassification due to out-of-vocabulary issues, thus limiting the efficacy of these analysers on code-switched data. Plus, obtaining a code-switched corpus annotated with sentiment labels is scarce, and investigations on sentiment analysis on code-switched social media data are lacking, particularly on Malaysian social media posts. We proposed MESocSentiment, a Malay-English social media corpus with sentiment labels constructed via a semi-automatic sentiment identification method to address this challenge. The framework leveraged existing language identifiers and sentiment analysers to annotate code-switched data. We collected 229,566 tweets containing #Malaysia between August 12, 2022, and May 15, 2023. Using our strategy, we identified 19714 code-switched posts containing Malay and English words from the collection. Our analysis showed that 78.23% of the corpus had neutral sentiments, while 16.32% were positive and 5.44% negative. Furthermore, the descriptive analysis of tweet length revealed a range spanning 43 words, with a mean of 8.92 and a standard deviation of 5.56 words. This comprehensive framework contributes to a deeper understanding of sentiment expression within code-switched social media data, particularly in the context of Malaysia's linguistic and cultural diversity. Additionally, the MESocSentiment corpus is published on GitHub for future research.

1. Introduction

Sentiment analysis or opinion mining is one of the tasks of Natural Language Processing (NLP) that is generally used to get the public's perception of specified topics. According to Liu [1], sentiment analysis or opinion mining is a research area that deals with people's opinions, sentiments, emotions, appraisals, and attitudes toward entities and their characteristics in text. Examples of entities include

[*] *Corresponding author.*
E-mail address: sjsflora@unimas.my

https://doi.org/10.37934/ard.123.1.198212

products, services, organisations, events, issues, and topics [1]. Sentiment analysis has been used in various cases, such as the public's perception of the COVID-19 vaccination program in Malaysia by Ariff *et al.,* [2] and in the Philippines by Co *et al.,* [3]. Meanwhile, the study by Sofian *et al.,* [4] utilised sentiment analysis to get the public's opinions on COVID-19 vaccine acceptance for children using Twitter data.

Sentiment analysis is also used to get customers' feedback about products or services, such as online food delivery services in Malaysia, as shown in the study by Samah *et al.,* [5]. Additionally, sentiment analysis is applied in the tourism industry, as shown in the study by Zaman *et al.,* [6], to find customers' sentiments on the place of interest (PoI) in Malaysia. Millions of users worldwide utilise social media platforms such as Facebook, Twitter, and YouTube, employing various languages in their interactions. Many of these users possess the ability to communicate in two or more languages, leading to instances of code-switching. This phenomenon becomes particularly pronounced on social media sites, where individuals seamlessly integrate multiple languages within their posts. In Malaysia, the two most prominently used languages are Malay and English, prompting a substantial number of users to employ both languages in their social media content.

According to Poplack [7], code-switching is the situation of mixing different languages in a communication. Code-switching presents challenges for the sentiment analysis task, such as variation of spelling, informal grammar forms, and scarcity of annotated data, according to the work by Aguero *et al.,* [8. Multilingual speakers usually use a mix of languages during informal communication, where the grammar rules are relaxed. The study by Srinivasan and Subalalitha [9] stated that the lenient grammar rules lead to various characteristics, including multiple variations of spellings of words and informal sentence structures. Hence, this informal communication can be challenging for sentiment analysis models trained using formal language. Spelling variations and informal sentence structure will cause challenges for sentiment analysis to give accurate sentiments.

Another challenge is the scarcity of annotated data for sentiment analysis of code-switched texts. Furthermore, according to the work by Khanuja *et al.,* [10], obtaining a code-switched corpus with sentiment labels is still scarce on sentiment analysis on code-switched social media data is lacking. This situation is particularly true on Malay-English code-switched sentiment corpus from Malaysian social media posts. Consequently, efforts to build classifiers that can handle code-switching are still challenging in terms of time and cost due to the unavailability of labelled code-switched corpus. Sentiment analysers built with models trained with monolingual data will cause misclassification due to out-of-vocabulary issues, thus limiting the efficacy of these analysers on code-switched data.

Therefore, we propose a framework to build a Malay-English Social Media Sentiment (MESocSentiment) corpus that can be used for sentiment analysis tasks. The challenge in building this framework is to design a strategy that uses existing classifiers to generate initial labels for the social media data. Therefore, the paper intends to investigate the following question: How to formulate a semi-automatic sentiment identification approach suitable for dealing with Malay-English code-switched data? This research question has brought to the objectives of this paper, which are:

i. To collect social media posts containing Malay-English tweets to build a code-switched corpus.
ii. To evaluate sentiment polarities for the code-switched corpus using a semi-automatic approach.

1.1 Related Works to Sentiment Analysis in Code-Switched Data

Code-switching generally involves using two or more languages in the same sentence. This linguistic phenomenon typically occurs in communications within bilingual and multilingual societies due to their fluency in two or more languages. The utilisation of code-switching in communication can be attributed to various reasons. Roslan *et al.,* [11] found that the top three reasons undergraduate students use Malay-English code-switched text on WhatsApp are habitual expressions, emphasising a point, and lack of vocabulary. The first reason is that students usually use discourse particles such as "kan" and "lah" in their communication [11]. Secondly, they also use code-switched texts, assuming that changing language during an argument can help strengthen the message. They also use code-switching when there is difficulty in explaining a matter in their first or second languages. Plus, they also change language when they are unable to find similar words in their first language. Meanwhile, Lubis *et al.,* [12] also found several reasons why late adolescents post code-switched posts on Facebook. One of the reasons is repetition. It is used to clarify or emphasise the message. An example is "Siap-siap kerja..prepare to work" post. In this example, the message suggests the user is ready to start working. The message is emphasised by repeating the same message in Bahasa Indonesia into English words.

The code-switched texts present several challenges for the sentiment analysis task. One of these challenges is the limited availability of annotated resources for code-switched data. Standardised datasets in code-switched languages are currently scarce besides datasets of a few language pairs used in shared tasks, as stated in Khanuja *et al.,* [10]. This is because there is a shortage of data and annotated resources for code-switched languages, even if one or both languages in the language pair are well-resourced [10]. Plus, the prevalence of different languages is varied, thus making the annotation process laborious and costly, as mentioned in Ranjan *et al.,* [13].

Another challenge in performing sentiment analysis on code-switched data is that code-switching can vary depending on individual preferences. Consequently, it becomes challenging to uncover underlying deep compositional semantics within code-switched text, as stated in Poria *et al.,* [14]. Code-switched text data have various characteristics due to their occurrence in informal communication. Variation of spellings occurs according to users' choice because there are no consistent spelling rules. Thus, various spellings of words have made code-switched texts hard to normalise for analysing sentiments. There is also no word order in code-switched texts because it depends on users to structure their own sentences [10], making sentiment analysis more complex. Besides these challenges, the presence of ambiguous words in code-switched data. In the study by Srivastava and Mayank [15], ambiguous words cannot be definitively classified into a specific language within a language pair without considering the context of the text. Thus, it is not easy to evaluate the annotations and proficiency of human annotators due to the characteristics of code-switched texts, as mentioned in the study by Srivastava and Mayank [16].

Several works on sentiment analysis have used code-switched data involving Malay and English languages. Kasmuri and Halizah [17] developed the MY-EN-CS corpus for sentiment analysis using entries from personal blogs. However, the corpus is not publicly available and did not use data from any social media sites. Instead of focusing on the Malay-English code-switched dataset, Romadhona *et al.,* [18] focused on Bahasa Rojak, a famous dialect in Malaysia that consists of Malay, English, and Chinese. The authors constructed the Bahasa Rojak Crawled Corpus (BRCC) using data augmentation techniques to pre-train language models for code-mixing phenomena. The author compiled the SentiBahasaRojak sentiment analysis dataset to evaluate the effectiveness of the Mixed XLM model developed by the authors. However, one of the SentiBahasaRojak dataset data sources was from financial and stock websites instead of fully using social media data. Kong *et al.,* [19] recently focused on sentiment analysis on a mixed Malay-English COVID-19 Twitter dataset. The dataset consisted of Malay and English and had three-class (positive, negative, or neutral) and two-class (positive and negative) sentiment labels. However, the dataset is not fully available for public use due to restrictions on access.

There are quite number of techniques that have been used in code-switching sentiment analysis including conventional machine learning techniques and deep learning models. Some of the machine learning techniques used in code-switching sentiment analysis are Support Vector Machine (SVM), Naïve Bayes, and Random Forest. After that, some of the used deep learning models in code-switching sentiment analysis include Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT). In Singh *et al.,* [20], SVM achieved the lowest accuracy among all used techniques and models for the sentiment classification task of collected Nepali social media data in both not concatenated and concatenated scenarios, with 0.684 and 0.714 values, respectively. The collected data in Singh *et al.,* [20] also included code-switched English-Nepali texts.

The Naïve Bayes algorithm was used by Jamatia *et al.,* [21] in two datasets. In the 10-fold cross-validation experiment, it achieved accuracy values of 42.4 and 45.7 in the English-Bengali and English-Hindi ICON2017 datasets, respectively [21]. Random Forest was also used in the same work for both datasets. The English-Bengali ICON 2017 dataset (10-Fold Cross-Validation) achieved the highest accuracy among machine learning algorithms, with 44.4 accuracy. The English-Hindi ICON 2017 dataset (10-Fold Cross-Validation) achieved 48.3 accuracy.

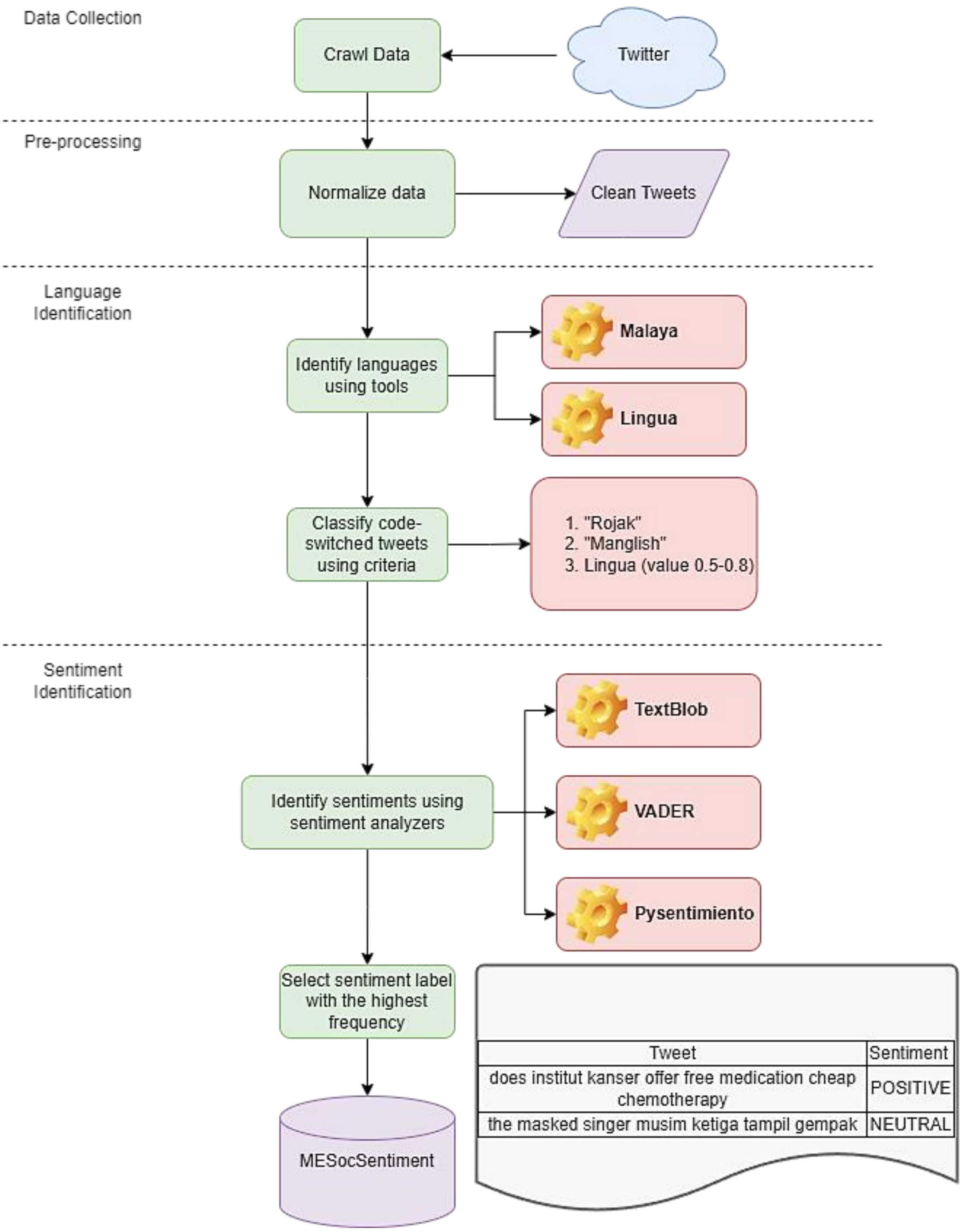
Some of the works for code-switching sentiment analysis also applied deep learning models in their works. In Thara and Prabaharan [22], the LSTM model with FastText word embedding has achieved 0.7381 accuracy compared to the LSTM model with Word2Vec with an accuracy of 0.6995. The result is for sentiment analysis for the code-mixed Malayalam-English dataset [22]. The authors also used CNN to analyse the sentiment of the same dataset. CNN model with FastText word embedding achieved 0.7455 accuracy compared to the CNN model with Word2Vec with an accuracy of 0.7329. A comparison of both results showed that CNN models achieved better accuracy than LSTM models in this work [22]. Younas *et al.,* [23] used the multilingual BERT model (mBERT) in the sentiment analysis of the code-mixed Roman Urdu-English dataset in their paper. In comparison to the XLM-R model, it achieved higher accuracy (0.65) without fine-tuning of hyper-parameters. Still, it had lower accuracy than XLM-R (0.71), with 0.69 accuracy, and with hyperparameter fine-tuning [23].

In Kasmuri and Halijah's [17] work, the authors used lexicons to identify code-switched sentences using Malay and English words from chosen blogs' entries. Sentences were annotated as either factual or opinion sentences. Romadhona *et al.,* [18] also developed a pre-trained model named "Mixed XLM" that can automatically tag the language of input tokens to process code-mixing input. The authors used the SentiBahasaRojak sentiment analysis dataset to evaluate the effectiveness of the model. The authors in Kong *et al.,* [19] collected 108,246 tweets from September to December 2021, with 67% in Malay, 27% in English, and the rest in Chinese and other languages. After that, 11,568 tweets were randomly chosen to be manually annotated to create the MyCovid-Senti dataset consisting of Malay and English.

Nevertheless, the three works on sentiment analysis for code-switched Malay-English data motivated the development of a computational framework to construct the MESocSentiment corpus. We are motivated to obtain this corpus because there is still a lack of publicly available corpus containing social media data in Malay-English. Plus, the creation of this corpus addresses one of the challenges for sentiment analysis of code-switched texts: the lack of annotated data and resources.

2. Methodology

The framework for constructing the MESocSentiment corpus has been created to answer the first and second research objectives. Acquiring the MESocSentiment corpus encompassed four steps, illustrated in Figure 1. The framework consisted of four stages: data collection, pre-processing, language identification, and sentiment identification. The subsequent subsections comprehensively described these steps as they were executed throughout the research.



Tweet	Sentiment
does institut kanser offer free medication cheap chemotherapy	POSITIVE
the masked singer musim ketiga tampil gempak	NEUTRAL

Fig. 1. A semi-automatic sentiment identification framework for MESocSentiment

2.1 Social Media Data Collection

Tweets were collected using Tweepy [24] library and Twitter API from 12 August 2022 until 15 May 2023. We used #Malaysia to crawl tweets, a similar strategy used by Zabha *et al.,* [25]. This specific keyword selection was made exclusively to accumulate tweets originating from Malaysia.

2.2 Pre-Processing

The newly acquired social media data requires normalisation to remove noise and stop words. Following are the steps taken to pre-process the data:

i. Removed tags and other invalid characters
ii. Converted the texts into lowercase
iii. Removed hashtags, mentions, and URL
iv. Removed numbers, punctuations, and stop words
v. Reduced whitespaces
vi. Removed sentences that only have one word or token
vii. Expanded English contractions to standardise text
viii. Removed duplicate sentences

2.3 Language Identification

After normalising the social media data, the next step was identifying the type of language used in the tweets. We acquired large amounts of tweets to create our code-switched corpus. Manually classifying whether each tweet belongs to Malay, English, or code-switching between the two would be laborious and costly. Thus, we utilised two language identification tools to classify the data. We selected Malaya [26] and Lingua [27] libraries because both tools support Malay and English identification.

We used Malaya's FastText model to identify the language of a sentence (post). In the current version, FastText was limited to identifying six languages (categories): Malay, English, Indonesian, Rojak, Manglish and Other. The model identified the language of the data by providing a probability value to one of the languages mentioned. Lingua is a natural language processing library in Python that provides language detection functionality using rule-based and statistical methods. Lingua can be set to detect specified language(s) supported by the library. In this research, it was set to detect Malay and English languages in tweets. It provided a label to an input text with confidence values from 0 to 1 for each specified language.

2.4 Sentiment Identification

Tweets under the 'rojak' and 'manglish' categories from the Malaya dataset and tweets ranging from 0.5, 0.6 and 0.7 groups from the Lingua dataset were classified as code-switched data. In this final step, code-switched data were labelled with negative, neutral, or positive sentiment polarities. We used three existing sentiment analysers, TextBlob [28], pysentimiento [29], and VADER [30], to generate pseudo-labels for the code-switched data. TextBlob is a Python library used to process textual data. The library is available for various natural language processing tasks, including sentiment analysis. TextBlob was chosen in this work because it is one of the famous tools used for sentiment identification and is easy to use. In Co *et al.,* [3], the collected data contained code-switched English-Tagalog (Taglish) tweets. There were two sets of datasets, one original and one translated. TextBlob, VADER, and Polyglot were used as existing sentiment analysers to measure the baseline performance of original and translated datasets. For the baseline performance of sentiment, TextBlob achieved an accuracy of 48.10% for the original dataset and 47.30% for the translated dataset [3].

The pysentimiento toolkit [aa] is trained for sentiment analysis in English, Spanish, Portugeuse, and Italian. This library was chosen because it is also trained in English tweets for the sentiment analysis task using the dataset from Rosenthal *et al.,* [31]. In Movahedi *et al.,* [32], pysentimiento was used to classify sentiments of English and Spanish tweets and Facebook posts collected by the authors. From the result, English posts had the highest negative posts, and Spanish posts had the fewest negative posts. Spanish posts also had higher neutral posts than English posts [32].

VADER is a rule-based model that is used for sentiment analysis. The library was chosen because VADER was made to focus on social media texts. The library is only widely used for sentiment analysis. Tho *et al.,* [33] used VADER and SentiWordNet models to classify code-mixed Javanese and Indonesian tweets' sentiments. Based on the performance, VADER was better than SentiWordNet overall. The results from each sentiment analyser were compared to get the final sentiment output. This voting approach provided an initial level of confidence in the final labels of the dataset. We expected ambiguous results due to the number of sentiment analysers used. The ambiguous case was when all sentiment analysers gave different sentiment labels for a tweet. For this kind of result, we removed the tweet from the dataset. The final dataset would be the MESocSentiment corpus.

3. Results

229, 566 tweets were collected using the keyword #Malaysia from 12 August 2022 until 15 May 2023. After applying the pre-processing step, we gained 138, 646 cleaned tweets. Subsequently, a language identification procedure was executed employing the Malaya and Lingua libraries on the tweets. Based on the outcomes, threshold values were selected to identify the code-switched tweets to be used in the subsequent step, sentiment identification.

3.1 Language Identification

We present the results of the language identification step based on the tool applied:

3.1.1 Malaya

For Malay language identification using the Malaya fastText model, the probability of the detected language was assigned to one of the supported languages. Table 1 shows some samples from the dataset for the language identification task using Malaya. All samples were code-switched tweets but were categorised into different categories by the library. In Tweet 2, it was categorised as Manglish. The structure of the sentence was in English but the phrase 'institut kanser' was used rather than 'cancer institute'. Tweet 5 also showed a code-switched Malay-English tweet categorised as Rojak by the tool. The sentence started with the first three words in English, and the following words were in Malay.

Table 1
Sample of language identification results using Malaya

Number	Tweets	Malay	English	Bahasa Indonesia	Manglish	Rojak
1	Pusat hiasan ikan near teluk kemang nice small aquarium lots colorful fish free get tickets popular things klook	0.534	0.0	0.0	0.0	0.0
2	Does institut kanser offer free medication cheap chemotherapy	0.0	0.0	0.0	0.596	0.0
3	Gurkhali machas bars rap titled suara rakyat	0.0	0.785	0.0	0.0	0.0
4	Jangan lupa ubah akta perumahan perolehi rumah dibina menghuni kaya buat untungan impose empty house tax like canada	0.0	0.0	0.774	0.0	0.0
5	The masked singer musim ketiga tampil gempak	0.0	0.0	0.0	0.0	0.849

3.1.2 Lingua

For Lingua, results for tweets were listed with confidence values between 0 and 1. For each tweet, the confidence value was rounded down to 3 decimal places to make it easier to analyse. For Lingua, tweets could be grouped into six groups based on the higher confidence values in each tweet. Each tweet received two confidence values, each for Malay and English languages. The groups were 1.0 group, 0.9 group, 0.8 group, 0.7 group, 0.6 group and 0.5 group.

In Table 2, all the samples shown were code-switched Malay-English sentences but had different confidence values in Malay and English. Tweet 7 contained Malay and English words but was given a 1.0 confidence value in Malay. This tweet fell under group 1.0. Tweet 2 was in the 0.9 group because its higher confidence value was 0.994 for English. The sentence structure was in English but used the Malay phrase ‘institut kanser’ instead of ‘cancer institute’. Tweet 3 was in group 0.7 because it had a 0.730 confidence value for Malay. Tweet 4 was in group 0.8 due to having a confidence value of 0.831 for English. It contained 10 English words and 4 Malay words. Tweet 5 had a confidence value of 0.686 for English. It had two English words followed by the Malay word ‘merdeka’. The word ‘merdeka’ was repeated three times consecutively. Tweet 6 was under group 0.5 because it had a confidence value of 0.571 for English. Malay words in Tweet 6 were used for the place and cuisines.

Table 2
Sample of language identification for Lingua

Number	Tweets	English	Malay
1	Pusat hiasan ikan near teluk kemang nice small aquarium lots colorful fish free get tickets popular things klook.	0.945	0.054
2	Does institut kanser offer free medication cheap chemotherapy.	0.994	0.005
3	Gurkhali machas bars rap titled suara rakyat.	0.269	0.730
4	Start new life circle baru vibe baru let flow meski kadang sad fell lonely.	0.831	0.168
5	Happy independence merdeka merdeka Merdeka.	0.686	0.313
6	Authentic malay cuisine kampung melayu subang delicious masak lemak smoke duck fried cat fish ulam more.	0.571	0.428
7	Inisiatif bagus tapi sepatutnya declare tegas saman tamat diskaun jangan ajar tunggu diskaun selesaikan saman hormat undang melindungi nyawa.	0	1.0
8	The masked singer musim ketiga tampil gempak.	0.091	0.908

After that, tweets under 'Rojak' and 'Manglish' for Malaya and tweets with confidence values of 0.5 to 0.8 for Lingua were combined to create a code-switched Malay-English dataset. The total number of tweets for this dataset was 20, 226 tweets.

3.2 Sentiment Identification

Table 3 shows the result of the voting approach in the sentiment identification stage. A voting approach using three existing sentiment analysers, TextBlob, VADER and Pysentimiento, was utilized in the sentiment identification stage. As a result of the voting approach, each tweet was labelled with either one of the sentiment categories or fell into the ambiguous category. The results showed that most tweets were labelled with neutral sentiment followed by positive and negative sentiment. The ambiguous category happened where all sentiment analysers gave different sentiment labels for a tweet. The ambiguous category was 'POSITIVE/NEGATIVE/NEUTRAL'. Ambiguous cases were removed to ensure the corpus only contains tweets from absolute categories which are positive, negative, and neutral.

Table 3
The result of the voting approach

Sentiment	Number of Tweets	Percentage (%)
Neutral	15422	76.23
Positive	3219	15.92
Negative	1073	5.31
Positive/negative/neutral	512	2.53

The final statistics for the MESocSentiment corpus are shown in Table 4. The MESocSentiment corpus was created after the removal of ambiguous tweets. Based on Table 4, the total number of tweets in MESocSentiment is 19,714 tweets. The corpus has tweets that have been mostly labelled as neutral, with 15,422 tweets, and negative tweets are the lowest, with 1,073 tweets. Table 5 shows some samples of tweets in the MESocSentiment corpus.

Table 4
The statistics of the MESocSentiment corpus in terms of sentiments

Sentiment	Number of Tweets	Percentage (%)
Neutral	15422	78.23
ositive	3219	16.33
Negative	1073	5.44

Table 5
Sentiment labels on random tweets from MESocSentiment after using the sentiment analysers

Number	Tweets	TextBlob	VADER	Pysentimiento	All
1	Does institut kanser offer free medication cheap chemotherapy.	Positive	Positive	Neutral	Positive
2	Telur supply tergugat ikan bilis too travesty injustice done holy dish nasi lemak.	Neutral	Negative	Negative	Negative
3	The masked singer musim ketiga tampil gempak.	Neutral	Neutral	Neutral	Neutral

Results from each sentiment analyser were compiled and compared for each tweet. The sentiment label with the highest count would be the final sentiment of the tweet. Tweet 1 was labelled POSITIVE by TextBlob and VADER but NEUTRAL by pysentimiento. Thus, the final sentiment of Tweet 1 is POSITIVE. Meanwhile, Tweet 2 was labelled as NEGATIVE sentiment by all sentiment analysers, making its final sentiment NEGATIVE. Tweet 3 was the same case as Tweet 2 but with the NEUTRAL sentiment label.

4. Discussions

The first research objective for this paper was to collect code-switched Malay-English tweets to build a code-switched corpus. The next research objective is to evaluate the sentiment polarities of the code-switched corpus using a semi-automatic approach. We developed the framework to construct the MESocSentiment corpus to achieve both research objectives. The first stage in the framework was data collection from social media. The result was 229 566 tweets collected by using the keyword #Malaysia from 12 August 2022 until 15 May 2023. The first research objective is to collect social media posts containing Malay-English tweets to build a code-switched corpus has been achieved in this stage. After that, 138 646 clean tweets were gained after pre-processing steps had been applied to the collected data in the preprocessing stage.

This is followed by the language identification stage, where Malaya and Lingua were chosen as language identification tools. The two tools were chosen because they support English and Malay language identification. After the identification, tweets under 'Rojak' and 'Manglish' for Malaya and tweets with confidence values of 0.5 to 0.8 for Lingua were combined, amounting to 20, 226 tweets. The dataset then underwent the sentiment identification stage. In this stage, the tweets were passed through three existing sentiment analysers: VADER, TextBlob, and pysentimiento. Results from these tools were compiled and compared by tweets. The MESocSentiment corpus was created after the ambiguous tweets were removed. This corpus contains 19, 714 tweets, with most being neutral tweets. The second research objective was achieved in this stage, where a semi-automatic approach was utilised to evaluate sentiment polarities of the code-switched corpus.

In language identification using Malaya's fastText model, the probability was only assigned to one category, as shown in the result in Table 1. The way the result was shown made it easier to choose code-switched Malay-English tweets by taking tweets that fall under 'manglish' and 'rojak' categories. However, based on this research definition of Malay-English code-switched sentences, there was also the probability of missing some of the other code-switching tweets. The definition of code-switched sentences in this research is that if only one foreign word comes from another language in a sentence, it is considered a code-switched sentence.

In language identification using Lingua, the presence of code-switching tweets, especially Malay-English tweets was more evident in the range of 0.5 to 0.8 even though there were also code-switching tweets in the higher range as well. Based on this research's current definition of code-switching, some tweets with a confidence value of 0.8 to 1.0 were code-switched tweets, but they occur less frequently. Besides, there were some presences of outlier tweets in other languages besides Malay and English, but they were still present due to the coding process. Using Malaya and Lingua for language identification reduced the need to find fluent speakers in Malay and English to classify the tweets. However, there was also some noise after the task due to the capability of both libraries. Three sentiment analysers (VADER, TextBlob, and pysentimiento) were used to identify tweets' sentiments via a voting approach. Three sentiment analysers were used to enhance the reliability of the sentiment identification process.

After the removal of ambiguous tweets, the current dataset is unbalanced and skewed more towards neutral categories. The dataset currently has 78.23% neutral tweets, 16.33% positive tweets, and 5.44% negative tweets. A descriptive analysis on the MESocSentiment corpus showed that the range of the length of tweets in the corpus was 43 with the lowest value being two words and the highest being 45 words. From the mean of 8.92 words, it could be seen that the average length of tweets in the dataset was around eight to nine words per tweet. Standard deviation was used to measure the variability of sentence length from the mean. The standard deviation of the dataset was 5.56 words. The result showed that most tweets in MESocSentiment were between three to 15 words per tweet. Currently, there are 15, 812 tweets that fall under this range, making 80.21% of the corpus. Tweets that fell under the POSITIVE/NEGATIVE/NEUTRAL category after the voting approach were considered ambiguous tweets. This was because each sentiment analyser gave different sentiment labels to the tweet. The ambiguous tweets were removed to create the MESocSentiment corpus. This was because the MESocSentiment corpus only contains tweets with one sentiment label. The sentiment label is the final sentiment from the voting approach using sentiment analysers.

Table 6 shows samples of true tweets from the MESocSentiment. It means the final sentiment label via voting approach is the same as when tweets are manually checked. For example, Tweet 1 is NEGATIVE because it shows some anger in the tweet. The context of the tweet is that an oath made by a group is useless due to bribery. Meanwhile, Tweet 2 is POSITIVE because it shows happiness in the sentence. The tweet's context is wishing a happy Independence Day to a neighbouring country. Lastly, Tweet 3 is neutral because it only states a statement about a group of leaders meeting the king.

Table 7 shows samples of false tweets from the MESocSentiment corpus. It means that the sentiment label given by the sentiment analysers is false when checked manually. For example, Tweet 1 was labelled as NEUTRAL instead of POSITIVE. This is because the tweet was about a beautiful village that has become a tourist destination for photography and camping. Tweet 2 was labelled as NEUTRAL instead of POSITIVE despite the sentence about wishing followers of the account happy Deepavali. Lastly, Tweet 3 should be positive sentiment instead of neutral sentiment. This is because the tweet was about an athlete who became a champion in a championship.

Table 6
Samples of true tweets

Number	Tweets	Sentiment (all)	Sentiment (true)
1	Their called sumpah worthless dna makan suap rasuah.	Negative	Negative
2	Happy birthday neighbour across causeway selamat Merdeka.	Positive	Positive
3	Gabungan leaders enter istana audience agong.	Neutral	Neutral

Table 7
Samples of false tweets

Number	Tweets	Sentiment (all)	Sentiment (true)
1	Antara tarikan penang kampung ni memang cantik sangat sampai tarikan bergambar camping kampung agong campsite penaga pulau pinang sumber Kampung agong.	Neutral	Positive
2	Selamat deepavali buat followers bikers ranger.	Neutral	Positive
3	Syabda raih gelar juara international series.	Neutral	Positive

Tweets in the MESocSentiment corpus were collected using the #Malaysia keyword to get Malaysian tweets. The collected tweets came from various topics, such as sports, politics, and art. Some tweets in sports categories were about users congratulating athletes on their achievements in tournaments. Some tweets related to politics were about the 2022 Malaysian general election, which was held on 19 November 2022. Art-related tweets, such as single releases and concerts, were also present. There were also tweets about festivities and events such as Eid, Christmas, Deepavali, Lunar New Year, and Independence Day. Compared to MESocSentiment, both datasets from Patwa *et al.,* [34] were in Spanish-English (Spanglish) and Hindi-English (Hinglish). For both datasets, every word in each tweet was annotated for the language identification task. Both datasets also kept some tokens and characters removed from the MESocSentiment. These tokens and characters included special characters, hashtags, mentions, and URL links.

From the result of each sentiment analysers and MESocSentiment in Table 8, the number of neutral tweets from TextBlob was the nearest to the MESocSentiment with a difference of 1, 117 tweets. TextBlob also had the nearest number of positive tweets to MESocSentiment with a difference of 954 tweets. However, pysentimiento had the nearest number of negative tweets to the MESocSentiment with a difference of 32 tweets. Thus, the result of TextBlob sentiment classification was the nearest to the MESocSentiment dataset.

Table 8
Sentiment classification result for each sentiment analysers and MESocSentiment

Sentiment	TextBlob	VADER	Pysentimiento	MESocSentiment
Neutral	14305	13115	16916	15422
Positive	4173	4899	1757	3219
Negative	1236	1700	1041	1073

The semi-automatic approach in the language identification and sentiment identification stage gave advantages to the framework. Malaya and Lingua used in the language identification task minimized the use of fluent Malay and English speakers in identifying tweets' languages. This reduced the time needed to identify the language of words in collected data. In addition, the sentiment analysers were leveraged to make sentiment identification faster than manual labelling. It also eliminated the need to find speakers fluent in Malay and English to label the dataset. This procedure also reduced the time taken to label tweets compared to manual sentiment labelling.

There were also a few limitations of using the semi-automatic approach in the framework of getting the MESocSentiment corpus. The use of tools in language identification made the process faster but there were still some noises in the dataset due to tools' limitations. In addition, false tweets are also present in the MESocSentiment corpus in the sentiment identification process. This situation was due to the limitations of sentiment analyser tools used in the voting approach. It showed that although multiple libraries were used, there was still a probability of having false tweets in the dataset. These also led to limitations of the MESocSentiment corpus. Firstly, the dataset is imbalanced because more than half of the dataset is neutral tweets. This may cause biased sentiment models. Then, there is also the presence of noises which can affect the performance of the corpus.

5. Conclusions and Future Works

In this paper, we proposed a semi-automatic framework for sentiment identification in Malay-English code-switched data, and we demonstrated its application using Twitter data. The framework comprises four key stages: data collection, data pre-processing, language identification, and sentiment classification. We collected 229,566 tweets from August 12, 2022, to May 15, 2023, using the Tweepy library and Twitter API, focusing on tweets containing the keyword #Malaysia to collect tweets from Malaysia. After applying pre-processing steps, the dataset was reduced to 138,646 tweets.

Subsequently, the dataset underwent language identification for Malay and English using the Malaya and Lingua libraries. We combined 'Rojak' and 'Manglish' tweets with tweets having confidence values between 0.5 to 0.8 from Lingua, resulting in a dataset of 20,226 tweets. To determine sentiment labels, a voting approach was employed, utilising three existing sentiment analyzers (TextBlob, VADER, pysentimiento). In addition to neutral, positive, and negative sentiments, we encountered ambiguous cases where each sentiment analyser provided different sentiment labels for the same tweet. Ambiguous tweets were excluded from the dataset, resulting in the creation of the MESocSentiment corpus, which contains 19,714 tweets labelled as neutral, positive, and negative sentiments. Notably, the MESocSentiment corpus is skewed toward the neutral category, accounting for 78.23% of the entire dataset. Descriptive analysis of the corpus reveals a mean tweet length of 8.92 words, a range of 43 words, and a standard deviation of 5.56 words.

In our future work, we are proposing to experiment with sentiment modelling based on machine learning on the code-switched MESocSentiment data. Furthermore, we have published the dataset on GitHub (https://github.com/afifahms/MESocSentiment) for open research opportunities.

Acknowledgement

The authors wish to thank Universiti Malaysia Sarawak for the financial support of this project.

References

[1] Liu, Bing. *Sentiment Analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2020. https://doi.org/10.1017/CBO9781139084789

[2] Ariff, Mohamed Imran Mohamed, Nurul Erina Shuhada Zubir, Azilawati Azizan Azilawati Azizan, Samsiah Ahmad, and Noreen Izza Arshad Noreen Izza Arshad. "Malaysian views on COVID-19 vaccination program: A sentiment analysis study using Twitter." *Bulletin of Electrical Engineering and Informatics* 13, no. 1 (2024): 436-443. https://doi.org/10.11591/eei.v13i1.6097

[3] Co, Nicole Allison, Maria Regina Justina Estuar, Hans Calvin Tan, Austin Sebastien Tan, Roland Abao, and Jelly Aureus. "Development of bilingual sentiment and emotion text classification models from covid-19 vaccination tweets in the Philippines." In *International Conference on Human-Computer Interaction,* p. 247-266. Cham: Springer International Publishing, 2022. https://doi.org/10.1007/978-3-031-05061-9_18

[4] Sofian, Muhammad Adam Sani Mohd, Norlina Mohd Sabri, Ummu Fatihah Mohd Bahrin, N. Hrishvanthika, and Norulhidayah Isa. "Sentiment analysis on acceptance of COVID-19 vaccine for children based on support vector machine." *Journal of Advanced Research in Applied Sciences and Engineering Technology* (2024): 252-270. https://doi.org/10.37934/araset.58.2.252270

[5] Samah, Khyrina Airin Fariza Abu, Nur Shahirah Jailani, Raseeda Hamzah, Raihah Aminuddin, Nor Afirdaus Zainal Abidin, and Lala Septem Riza. "Aspect-based classification and visualization of Twitter sentiment analysis towards online food delivery services in Malaysia." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 37, no. 1 (2024): 139-150. https://doi.org/10.37934/araset.37.1.139150

[6] Zaman, Qiryn Adriana Kharul, Wan Nur Syahidah Wan Yusoff, and Qistina Batrisyia Azman Shah. "Sentiment analysis on the place of interest in Malaysia." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 43, no. 1 (2025): 54-65. https://doi.org/10.37934/araset.43.1.5465

[7] Poplack, Shana. 2015. "Code switching: Linguistic." In *International Encyclopedia of the Social and Behavioral Sciences*, *2nd Edition*, edited by James D. Wright, 918-925. Oxford: Elsevier. http://dx.doi.org/10.1016/B978-0-08-097086-8.53004-9

[8] Agüero-Torales, Marvin M., José I. Abreu Salas, and Antonio G. López-Herrera. "Deep learning and multilingual sentiment analysis on social media data: An overview." *Applied Soft Computing* 107 (2021): 107373. https://doi.org/10.1016/j.asoc.2021.107373

[9] Srinivasan, R., and C. N. Subalalitha. "Sentimental analysis from imbalanced code-mixed data using machine learning approaches." *Distributed and Parallel Databases* (2021): 1-16. https://doi.org/10.1007/s10619-021-07331-4

[10] Khanuja, Simran, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. "GLUECoS: An evaluation benchmark for code-switched NLP." *arXiv preprint arXiv:2004.12376* (2020). https://doi.org/10.48550/arXiv.2004.12376

[11] Roslan, Adlin Nadhirah Mohd, Malissa Maria Mahmud, and Othman Ismail. "Why code-switch on WhatsApp? A quantitative analysis of types and influences of code-switching." *Asian Social Science* 17, no. 10 (2021): 43-52. https://doi.org/10.5539/ass.v17n10p43

[12] Lubis, Indah Sari, Satyawati Surya, and Adinda Usin Muka. "The use of code-switching among the late adolescents in social media Facebook." *CaLLs (Journal of Culture, Arts, Literature, and Linguistics)* 3, no. 2 (2017): 83-96. http://dx.doi.org/10.30872/calls.v3i2.817

[13] Ranjan, Sudhanshu, Dheeraj Mekala, and Jingbo Shang. "Progressive sentiment analysis for code-switched text data." In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1155-1167. 2022. http://dx.doi.org/10.18653/v1/2022.findings-emnlp.82

[14] Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research." *IEEE Transactions on Affective Computing* (2020). https://doi.org/10.1109/TAFFC.2020.3038167

[15] Srivastava, Vivek, and Mayank Singh. "IIT Gandhinagar at SemEval-2020 task 9: Code-mixed sentiment classification using candidate sentence generation and selection." In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, p. 1259-1264. 2020. http://dx.doi.org/10.18653/v1/2020.semeval-1.168

[16] Srivastava, Vivek, and Mayank Singh. "Challenges and limitations with the metrics measuring the complexity of code-mixed text." In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, p. 6-14. 2021. http://dx.doi.org/10.18653/v1/2021.calcs-1.2

[17] Kasmuri, Emaliana, and Halizah Basiron. "Building a Malay-English code-switching subjectivity corpus for sentiment analysis." *Int. J. Advance Soft Compu. Appl* 11, no. 1 (2019).

[18] Romadhona, Nanda Putri, Sin-En Lu, Bo-Han Lu, and Richard Tzong-Han Tsai. "BRCC and SentiBahasaRojak: The first bahasa rojak corpus for pretraining and sentiment analysis dataset." In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 4418-4428. 2022.

[19] Kong, Jeffery TH, Filbert H. Juwono, Ik Ying Ngu, I. Gde Dharma Nugraha, Yan Maraden, and W. K. Wong. "A Mixed Malay–English language COVID-19 Twitter dataset: A sentiment analysis." *Big Data and Cognitive Computing* 7, no. 2 (2023): 61. https://doi.org/10.3390/bdcc7020061

[20] Singh, Oyesh Mann, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. "Aspect based abusive sentiment detection in Nepali social media texts." In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, p. 301-308. IEEE, 2020. https://doi.org/10.1109/ASONAM49781.2020.9381292

[21] Jamatia, Anupam, Steve Durairaj Swamy, Björn Gambäck, Amitava Das, and Swapan Debbarma. "Deep learning based sentiment analysis in a code-mixed English-Hindi and English-Bengali social media corpus." *International Journal on Artificial Intelligence Tools* 29, no. 05 (2020): 2050014. https://doi.org/10.1142/S0218213020500141

[22] Thara, S., and Prabaharan Poornachandran. "Social media text analytics of Malayalam–English code-mixed using deep learning." *Journal of Big Data* 9, no. 1 (2022): 45. https://doi.org/10.1186/s40537-022-00594-3

[23] Younas, Aqsa, Raheela Nasim, Saqib Ali, Guojun Wang, and Fang Qi. "Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches." In *2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE)*, p. 66-71. IEEE, 2020. https://doi.org/10.1109/CSE50738.2020.00017

[24] Roesslein, Joshua. *Tweepy Documentation*. Tweepy, 2023.

[25] Zabha, Nur Imanina, Zakiah Ayop, Syarulnaziah Anawar, Erman Hamid, and Zaheera Zainal Abidin. "Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach." *International Journal of Advanced Computer Science and Applications* 10, no. 1 (2019). https://dx.doi.org/10.14569/IJACSA.2019.0100146

[26] Zolkepli, Husein. *Malaya: Natural-Language-Toolkit Library for Bahasa Malaysia, Powered by PyTorch*. 2018. GitHub repository.

[27] Gehweiler, Christoph, and Oleg Lobachev. "Classification of intent in moderating online discussions: An empirical evaluation." *Decision Analytics Journal* 10 (2024): 100418. https://doi.org/10.1016/j.dajour.2024.100418

[28] Loria, Steven. *TextBlob Documentation: Release 0.18.0.post0.* 2024.

[29] Pérez, Juan Manuel, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. "pysentimiento: A Python toolkit for opinion mining and social NLP Tasks." arXiv (2021). https://doi.org/10.48550/arXiv.2106.09462

[30] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1, pp. 216-225. 2014. https://doi.org/10.1609/icwsm.v8i1.14550

[31] Rosenthal, Sara, Noura Farra, and Preslav Nakov. "SemEval-2017 task 4: Sentiment analysis in Twitter." *arXiv preprint arXiv:1912.00741* (2019). https://doi.org/10.48550/arXiv.1912.00741

[32] Movahedi Nia, Zahra, Nicola Bragazzi, Ali Asgary, James Orbinski, Jianhong Wu, and Jude Kong. "Mpox panic, infodemic, and stigmatization of the two-spirit, lesbian, gay, bisexual, transgender, queer or questioning, intersex, asexual community: Geospatial analysis, topic modeling, and sentiment analysis of a large, multilingual social media database." *Journal of Medical Internet Research* 25 (2023): e45108. https://doi.org/10.2196/45108

[33] Tho, C., Y. Heryadi, L. Lukas, and A. Wibowo. "Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach." In *Journal of Physics: Conference Series* 1869, no. 1, p. 012084. 2021. https://doi.org/10.1088/1742-6596/1869/1/012084

[34] Patwa, Parth, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas Pykl, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. "Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets." *arXiv preprint arXiv:2008.04277* (2020). http://dx.doi.org/10.18653/v1/2020.semeval-1.100