An Enhanced Framework Using XLM-R with Optimized TF-IDF and Positional Encoding for Intra-Sentential Code Mixing Malay-English Sentiment Analysis

Surendran Selvaraju[1]*, Nilam Nur Amir Sjarif[2], Nurulhuda Firdaus Mohd Azmi[3],
Wan Noor Hamiza Wan Ali[4], Norshaliza Kamaruddin[5]
Faculty of Artificial Intelligence, University Technology Malaysia, Kuala Lumpur, Malaysia

Abstract*—The increasing use of online platforms, especially in social media, has led to a rapid growth of user-generated content that frequently exhibits intra-sentential code mixing between Malay and English language. Sentiment analysis remains challenging due to linguistic heterogeneity, frequent language switching, non-standard syntax and limited availability of adequate representations for code mixing text. Although multilingual contextual embedding models such as Cross-lingual Language Model (XLM-R) provide good semantic representations, but there are still challenges in capturing fine-grained sentiment cues in intra-sentential code mixing text when used directly. This study proposed an enhanced feature extraction framework for intra-sentential code mixing Malay-English. The framework first constructs TF-IDF weighting based on trigrams and followed by lexicon-guided filtering to select trigrams that contain sentiment-relevant words. Contextual embeddings are then extracted using XLM-R and further refined through Term Frequency–Inverse Document Frequency (TF-IDF) weighting and positional encoding to preserve structural information. The dataset derived from the MESocSentiment corpus with total of 4,292. The experimental results show that the proposed framework achieves an accuracy of 0.896 and an F1-score of 0.932, where it outperforms traditional sparse feature representations and multilingual contextual embedding baselines. Notably, the framework demonstrates a high recall of 0.954, indicating strong sensitivity in identifying sentiment-bearing instances across diverse social media code mixing expressions. Further analysis reveals that the integration of informative trigram filtering, XLM-R based contextual embedding, TF-IDF weighting, positional encoding, and sentiment polarity scoring enhances the representation of sentiment cues in short and informal social media text. Overall, the results suggest that the proposed feature extraction framework enhances the representation quality of sentiment analysis for code mixing Malay–English in social media.

Keywords*—Sentiment analysis; code mixing; feature extraction; contextual embeddings; XML-R; TF-IDF; positional encoding

I. Introduction

Social media platforms generate large volumes of user-generated content that reflect public opinions, experiences and attitudes [2,3,4,5]. In multilingual societies, such as Malaysia, this content frequently exhibits intra-sentential code mixing between Malay and English [4]. This phenomenon, commonly referred to as Bahasa Rojak, introduces substantial linguistic complexity due to rapid language switching, syntactic irregularities, informal expressions and frequent use of slang and out-of-vocabulary terms (OOV) [6,7]. Sentiment analysis of code mixing text remains a challenging task, as most of the existing approaches are designed for monolingual data and struggle to generalize across mixed linguistic structures [6,7,8,10]. While recent multilingual contextual embedding models such as Cross-lingual Language Model (XLM) have demonstrated good performance in multilingual natural language processing tasks, their direct application to code mixing sentiment analysis often yields suboptimal representations [11,12,13]. This limitation arises because contextual embedding alone may overlook sentiment-specific cues, negation effects, and local contextual dependencies that are critical for accurate sentiment interpretation in code mixing text [14,15,47,48,49].

Additionally, traditional sentiment analysis approaches based on sparse feature representations such as bag-of-words (BOW) and TF-IDF remain computationally efficient and effective in some settings. However, this approach focuses more on model-centric improvements and does not address the limitations of representation level in intra-sentential code mixing, where they are limited in capturing contextual semantics and short-range word order information. The studies, including mBERT and XLM-R, have improved cross-lingual representation learning and enabled sentiment analysis across languages. Nevertheless, contextual embedding alone may not sufficiently capture sentiment-relevant patterns in code mixing social media text where sentiment cues are distributed across local word combinations and contextual structures.

To address these gaps, this study evaluates the proposed feature extraction framework that introduces XLM-R based contextual embeddings with informative trigram filtering, TF-IDF weighting, positional encoding and sentiment polarity scoring. The framework is designed to enhance the representation of sentiment cues in intra-sentential code mixing Malay–English text by combining contextual semantics with local lexical patterns and polarity information. This study presents an empirical evaluation of the proposed feature extraction framework on code mixing Malay–English in social media dataset. This dataset derived from the MESocSentiment corpus. The dataset reflects realistic online communication

*Corresponding author.

patterns which include informal language usage, slang, short sentence length and diverse sentiment expressions. By focusing exclusively on this dataset, the evaluation examines the framework's performance and robustness in a realistic social media setting.

The contributions of this study are threefold. First, it provides a detailed evaluation of the feature extraction framework for code mixing Malay–English in a social media dataset. Second, it presents a comparative analysis against traditional sparse feature-based techniques and multilingual contextual embedding baselines, highlighting the benefits of integrating sentiment polarity score components with contextual representations. Third, it offers an error analysis that identifies key challenges in sentiment analysis of code mixing social media text, which includes mixed sentiment cues, informal expressions and pragmatic language use.

The remainder of this study is organized as follows: Section II reviews the related work in sentiment analysis for code mixing, while Section III presents the proposed feature extraction framework. Section IV describes the dataset and experimental setup. Section V presents the evaluation results and comparative analysis. Section VI concludes the study and outlines directions for future work.

II. Related Works

Sentiment analysis of code mixing text has received increasing research attention due to the widespread use of multilingual communication on social media and online platforms. Code mixing refers to the use of two or more languages within a single utterance, resulting in informal, non-standard text structures that pose significant challenges for natural language processing tasks. Early studies in this domain primarily focused on understanding the linguistic characteristics of code mixing data, particularly in high-resource language pairs, such as Hindi–English, which dominate the existing literature.

Initial approaches to code mixing sentiment analysis relied on traditional machine learning techniques and surface-level features including n-grams, TF-IDF and lexicon-based representations [9,16,17]. Choudhary in [25] introduced the Sentiment Analysis of Code mixing Text (SACMT) framework, which employed contrastive learning using Siamese neural networks to map code mixing and monolingual sentences into a shared sentiment space. While this approach demonstrated that sentiment similarity could be preserved across languages, it relied on skip-gram embeddings and required large auxiliary English datasets to achieve satisfactory performance, limiting its applicability to low-resource settings.

Subsequent research explored deep learning architectures to better capture contextual and sequential information in code mixing text. Lal *et al*. [26] proposed a hybrid model that combined sub-word embeddings with a dual Bi-LSTM encoder architecture, enabling the model to capture both sentence-level sentiment and sentiment-bearing sub-word units. Similarly, Yadav *et al*., in [27], employed Bi-LSTM models with sequential mirroring to enhance contextual learning in longer code mixing sentences. Although these neural approaches improved sentiment classification performance, they remained dependent on static or weakly contextualized embeddings which struggled to fully model the dynamic nature of intra-sentential code mixing.

To address the limitations of static embeddings, later studies shifted towards contextual embedding models such as BERT and XLM [41]. Contextual embeddings capture word meaning based on surrounding context, making them more suitable for code mixing text [11,12,13]. Researchers demonstrated that multilingual variants such as mBERT and XLM-R can learn shared semantic representations across languages, yielding improved sentiment analysis performance in multilingual and low-resource scenarios [11,10,28,29]. Studies by Tang *et al.* [30] and Sabri *et al*. [31] showed the effectiveness of multilingual word embedding for code mixing sentiment classification, although performance varied across sentiment categories and domains.

Recent work in monolingual sentiment analysis has focused on enhancing transformer-based models through hybrid and feature-augmented approaches. Jin *et al*. [32] demonstrated that incorporating TF-IDF weighting into contextual embeddings can improve sentiment discrimination by emphasizing statistically salient words in English text. Similarly, Mutinda *et al*. [33] proposed a hybrid representation that combines n-gram features, sentiment lexicons, and word embedding to enhance sentiment classification performance. Although these studies are not designed for code mixing data, they highlight the potential benefits of enriching contextual embeddings with explicit statistical and sentiment-based features. Liapis *et al*. [37] introduced distributional emotion embeddings that explicitly integrate emotional information into the embedding space, showing that enriching semantic representations with structured affective features leads to improved sentiment-related tasks. Similarly, language tagging and part-of-speech augmentation techniques proposed by Patil *et al*. [34], Tawakane *et al*. [35] and Patwardhan *et al*. [36], showed that explicit linguistic annotations can improve BERT-based model performance for code mixing sentiment analysis. However, these approaches often require extensive preprocessing, language-specific annotations, or large computational resources.

Research on Malay–English code mixing sentiment analysis remains comparatively limited. Kasmuri and Basiron [38] highlighted the scarcity of annotated Malay–English code mixing datasets and proposed the MY-EN-CS corpus to support future research. Lexicon-based methods explored by Zabha *et al*. [39] attempt to address data scarcity by constructing bilingual sentiment lexicons. However, these approaches were found to be ineffective in handling slang, abbreviations, and informal expressions commonly found in social media text. Fuady and Ibrahim [40] proposed a multilingual embedding approach using Word2Vec trained on Wikipedia corpora but their method struggled with out-of-vocabulary words and morphological variations.

More recent studies in code mixing Malay-English work have focused on multilingual transformer models such as mBERT and XLM-R [11,10,42]. Romadhona et al. [11] introduced the Bahasa Rojak Crawled Corpus (BRCC) and the SentiBahasaRojak dataset to support multilingual pretraining and evaluation of sentiment analysis models on Malay–English

code mixing text, including adaptations of XLM-R for mixed-language inputs. While their work improved cross-lingual semantic representation, sentiment polarity discrimination remained inconsistent, particularly between positive and negative classes in informal social media text. Kong et al. [10] constructed a large scale code mixing Malay–English COVID-19 Twitter dataset and evaluated sentiment analysis using BPE-enhanced mBERT and CNN-based models to address low-resource and noisy text challenges. Despite performance gains from subword modeling and contextual embeddings, the approach remains largely sentiment agnostic at the representation level. Suhaimin et al. [42] proposed MSSThreD, a multitask framework combining sentiment analysis and sarcasm detection using Bi-LSTM and GRU architecture with engineered linguistic features, demonstrating improved robustness to sarcasm-induced polarity reversal. While these approaches show promise, they continue to face challenges in accurately modeling sentiment polarity in intra-sentential code mixing text, particularly for positive and negative sentiment classes. Overall, existing research highlights that although contextual embeddings significantly improve semantic representation, they often lack explicit mechanisms to encode sentiment polarity, negation effects and local contextual structure.

In summary, prior work demonstrates that effective sentiment analysis of code mixing requires more than model-centric improvements. There remains a clear need for a principled feature extraction framework that explicitly integrates contextual semantics, local trigram structure, positional information and sentiment polarity, particularly for low-resource language pairs such as Malay–English. This gap motivates the proposed framework in this study.

III. Proposed Feature Extraction

The proposed feature extraction framework aims to enhance contextual embeddings for sentiment analysis by explicitly modeling local context, structural information and sentiment polarity scores. Given an input code mixing sentence, the framework produces sentiment-enriched trigram embeddings that serve as input representations for downstream classification models. Fig. 1 illustrates the overall workflow of the proposed framework.

The proposed model is detailed in Fig. 2 and Fig. 3. Based on both figures, the model is composed of five key modules, which are XLM-R embeddings, Informative trigrams, TF-IDF weighting, Positional Encoding (PE), and Sentiment Polarity Scores. Each module is strategically designed to enhance the quality of feature representations for sentiment analysis in intra-sentential code mixing Malay text.

A. Informative Trigram Generation

First, the preprocessed text is refined using an informative trigrams extraction mechanism, which selectively identifies only the most sentiment-relevant trigrams for downstream modelling. For a given code mixing sentence $S_{CodeMixed} = \{w_1, w_2, \ldots, w_n\}$, trigrams are first generated by sliding a three-word window across the sequence, forming candidates $T_i = \{w_i, w_{i+1}, w_{i+2}\}$ for all $i = 1$ to $W_n - 2$. Instead of retaining all possible trigrams, the proposed method filters them using a code mixing sentiment lexicon [43,44,45,46]. The process begins by filtering and retaining trigrams with at least one of its component words belonging to the lexicon. This filtering step discards trigrams with no sentiment bearing cues which significantly reduce the noise commonly found in code mixing social media text.

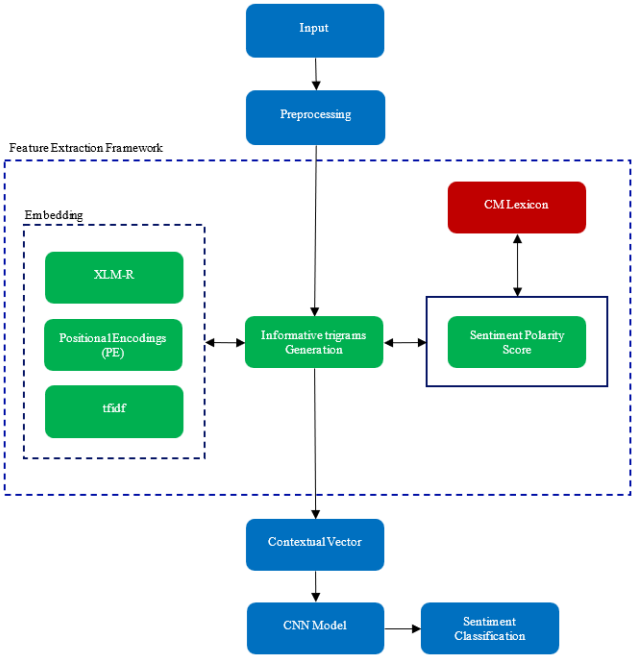

Fig. 1. Proposed feature extraction framework.

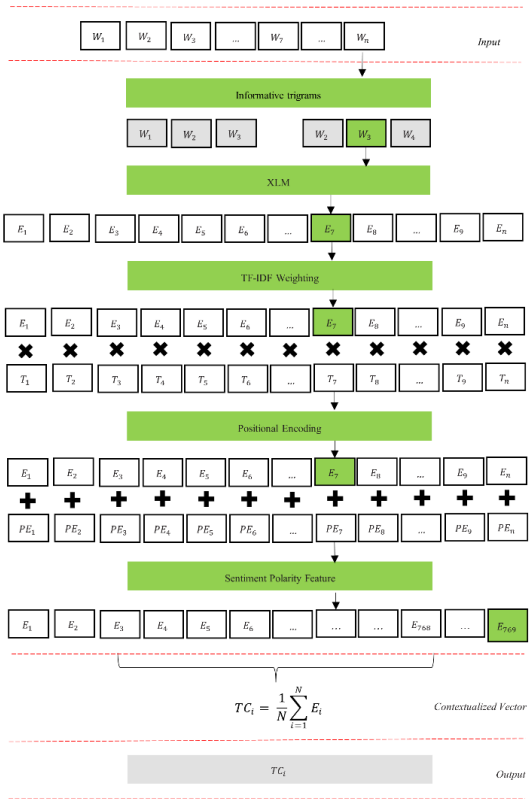

Fig. 2. Architecture of the proposed feature extraction.

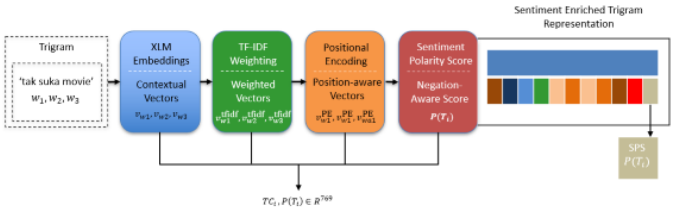

Fig. 3. Illustration of the flow process in code mixing Malay-English.

Each retained trigram is then scored using a combined metric incorporating TF-IDF importance and sentiment polarity score. The TF-IDF score captures the semantic weight of a trigram by averaging the TF-IDF values of its three component words, computed, as in Eq. (1):

$$TFIDF(T_i) = \frac{1}{3}\sum_{j=1}^{3} tfidf(w_{i+j-1}) \quad (1)$$

where, $tfidf(w)$ is the precomputed TF-IDF value for word w obtained from the dataset. Higher TF-IDF scores indicate that the trigram contains more informative and rich content terms. In parallel, the polarity score is derived using sentiment lexicon membership. Each word contributes $+1$ if it is labelled positive in the lexicon, -1 if it is labelled as negative in the lexicon and 0 otherwise. The polarity score for the trigram is then normalized, as shown in Eq. (2), to maintain a bounded range ensuring that $P(T_i) \in [-1, +1]$.

$$P(T_i) = \frac{1}{3}\sum_{j=1}^{3} polarity(w_{i+j-1}) \quad (2)$$

Normalization prevents trigrams with multiple sentiment words from dominating solely due to word count and provides a fair comparative scale across all candidates. The final informativeness score for ranking trigrams is computed by combining these two signals through a simple additive model, as shown in Eq. (3):

$$Score(T_i) = TFIDF(T_i) + P(T_i) \quad (3)$$

$Score(T_i)$ favors trigrams that are both semantically important and sentiment-heavy. Using this score, the algorithm selects the top K highest scoring trigrams. The use of a fixed top K selection ensures a bounded and comparable representation across sentences of varying lengths while suppressing weak or redundant sentiment cues. In this study, K is set to 20 to provide sufficient contextual coverage for sentiment expression without introducing excessive redundancy. To ensure that important sentiment expressions near the end of the sentence are not overlooked, especially in informal user-generated text, the method includes a tail coverage strategy in which the last few trigrams (typically two) are preserved regardless of their computed score. These tail trigrams are then merged with the top K set, deduplicated and sorted based on their original positional index within the sentence to maintain natural linguistic flow. Through this refined extraction process, the model effectively captures the most meaningful and sentiment-relevant textual segments, yielding a robust and contextually rich trigram representation tailored for sentiment analysis of intra-sentential code mixing Malay-English text. These selected informative trigrams provide the textual units on which contextual embeddings are subsequently extracted.

B. XLM-R Word Embedding

The proposed framework employs the Cross-lingual Language Model (XLM-R) to generate contextualized embeddings for intra-sentential Malay–English code mixing text. XLM-R is a transformer-based multilingual model pretrained on large scale parallel corpora that enables words from different languages to be represented within a shared semantic space. Given a code mixing sentence $S = \{w_1, w_2, \ldots, w_n\}$, each token w_i is first tokenized into subword units and mapped to a contextual embedding $w_{vn} \in R^d$ using XLM-R, where d denotes the embedding dimensionality. Unlike static embeddings, these representations are dynamically conditioned on the surrounding context, allowing XLM-R to capture semantic variation across Malay and English words within the same sentence. This property is essential for code mixing text where sentiment expressions frequently span multiple languages.

To preserve local contextual semantics, word-level embeddings extracted from XLM-R are aggregated at the trigram level. For a trigram $t_j = \{w_j, w_{j+1}, w_{j+2}\}$, the corresponding contextual representation is computed by aggregating the embeddings of its constituent words. However, as observed in this study, XLM-R embeddings primarily encode semantic and syntactic information and do not explicitly model sentiment polarity or structural importance. Therefore, the resulting trigram representations serve as a semantic foundation that is subsequently refined through TF-IDF weighting, positional encoding and sentiment polarity score integration to ensure that the final embedding is both contextually rich and sentiment aware.

C. TF-IDF Weighting

Following the extraction of word-level contextual embeddings using XLM-R, the proposed framework applies Term Frequency–Inverse Document Frequency (TF-IDF) weighting to emphasize sentiment-relevant words in code mixing text. TF-IDF is used to assign higher importance to words that occur frequently within a sentence but less frequently across the corpus, thereby reducing the influence of common and non-sentiment relevant words [19,20,21]. For a word w_n occurring in a sentence S within a document collection D, the TF-IDF weight is defined as Eq. (4):

$$TF - IDF(w_n) = TF(w_n, S) \times IDF(w_n) \quad (4)$$

where, the term frequency $TF(w_n, S)$ is given by Eq. (5):

$$TF(w_n, S) = \frac{f(w_n, S)}{|S|} \quad (5)$$

with $f(w_n, S)$ denoting the number of occurrences of w_n in sentence S, and $|S|$ representing the total number of tokens in S. The inverse document frequency is computed, as shown in Eq. (6):

$$IDF(w_n) = log\left(\frac{|D|}{|\{S_j \in D : w_n \in S_j\}|}\right) \quad (6)$$

where, $|D|$ denotes the total number of sentences in the corpus and $|\{S_j \in D: w_n \in S_j\}|$ represents the number of sentences containing the word w_n. Each contextual embedding v_w generated by XLM-R is then weighed using its corresponding TF-IDF score, as in Eq. (7):

$$v_w^{tfidf} = TF - IDF(w_n) \cdot v_w \tag{7}$$

The resulting weighted embeddings v_w^{tfidf} are subsequently used in the trigram construction stage of the framework. By incorporating TF-IDF weighting after contextual embedding extraction, the framework ensures that statistically significant and sentiment-relevant words contribute more strongly to the final representation, which is particularly important in intra-sentential code mixing text where frequent function words and language-switching artifacts may otherwise dominate the embedding space. However, weighing alone does not preserve word order within local contexts, which is critical for sentiment interpretation.

D. Positional Encoding

Although TF-IDF weighting emphasizes sentiment-relevant words, weighted embeddings alone do not explicitly preserve the relative position of words within local structures such as trigrams. In intra-sentential code mixing text, word order plays a critical role in sentiment interpretation, particularly in the presence of modifiers and negation [22]. For example, reversing the order of sentiment-bearing words within a short phrase may alter the overall sentiment polarity. To address this limitation, the proposed framework incorporates positional encoding to explicitly encode word order information within trigram-level representations.

Let w_n denote the i-th word in a sentence and $v_{wn}^{tfidf} \in R^d$ represent its TF-IDF-weighted contextual embedding. A positional encoding vector $PE_i \in R^d$ is assigned to each word based on its relative position within the trigram. Following the formulation in Eq. (8) adopted from Ali et al. [1], the positional encoding is defined using sinusoidal functions.

$$p_i^{(2k)} = \sin\left(\frac{i}{10000^{\frac{2k}{d}}}\right), \quad p_i^{(1)} = cos\left(\frac{i}{10000^{2k/d}}\right) \tag{8}$$

where, k denotes the dimension index and d is the embedding dimensionality. The position-aware embedding is then obtained by element-wise addition, as shown in Eq. (9):

$$v_{w1}^{PE} = v_w^{tfidf} + PE_i \tag{9}$$

The position aware embeddings v_{wn}^{PE} are subsequently aggregated at the trigram level to form structure preserving trigram representations, as depicted in Eq. (10):

$$TC_i = \frac{1}{3}\sum_{n=j}^{j+2} v_{wn}^{PE} \tag{10}$$

By incorporating positional encoding after TF-IDF weighting, the framework ensures that the final trigram representations retain both the statistical importance of words and their relative order within local context windows. This explicit encoding of positional information enhances the model's ability to capture sentiment shifts caused by word ordering, which is particularly important in code mixing text where linguistic structure is often irregular and informal. To further enhance sentiment sensitivity beyond structural encoding, explicit sentiment polarity information is incorporated.

E. Sentiment Polarity Score

To enhance the sentiment sensitivity of the feature representation, the proposed framework incorporates a Sentiment Polarity Score (SPS) derived from a code mixing sentiment lexicon and a negation scope-aware mechanism. Unlike conventional feature extraction techniques, this approach embeds the sentiment polarity score directly into the trigram-level contextual representation. This allows affective information to be explicitly encoded within the embedding space. This design is particularly suited to intra-sentential Malay–English code mixing text where sentiment-bearing words from different languages may co-occur and negation can substantially alter sentiment orientation.

For each informative trigram $T_i = \{w_1, w_2, w_3\}$, sentiment polarity is computed based on the lexicon membership of its constituent words. Each word w_n contributes a polarity value defined as in Eq. (11):

$$\text{polarity}(w_j) = \begin{cases} +1 & \text{if } w_j \in \text{positive polarity} \\ -1 & \text{if } w_j \in \text{negative polarity} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

To account for negation effects, a dynamic negation scope with a fixed window size of three tokens is applied. When a negation word (e.g., *tidak*, *not*, *bukan*) is encountered within the trigram, the polarity of the subsequent two tokens is inverted. This mechanism is implemented using a negation window counter that triggers polarity inversion for affected words. The negation-adjusted polarity values are then used to compute the overall sentiment polarity score of the trigram, as in Eq. (12):

$$P(T_i) = \frac{1}{3}\sum_{j=1}^{3} negated\_polarity(w_n) \tag{12}$$

where, $P(T_i) \in [-1, +1]$ represents a normalized polarity score capturing both sentiment intensity and directional shifts induced by negation. The resulting scalar polarity value is concatenated with the trigram's contextual embedding TC_i, producing a sentiment-enriched trigram representation, as in Eq. (13):

$$TC_i = [TC_i \parallel P(T_i)] \tag{13}$$

where, $TC_i \in R^{769}$ combines semantic and sentiment dimensions. By integrating sentiment polarity and negation scope directly into the embedding layer, the proposed framework enables more effective differentiation between trigrams that are semantically similar but sentimentally opposite [23,24]. This is an essential capability for sentiment analysis in code mixing contexts.

F. Final Trigram Representation

The final output of the proposed feature extraction framework is a sentiment enriched trigram representation that integrates semantic, structural and affective information. For each informative trigram, contextual embeddings extracted using XLM-R are refined through TF-IDF weighting to emphasize sentiment relevant words, augmented with positional encoding to preserve local word order and then further enriched by the integration of a negation aware sentiment polarity score (SPS), as illustrated in Fig. 2. These components are combined through vector concatenation to form a unified trigram representation that captures both contextual meaning and sentiment orientation. At the sentence level, the proposed framework represents each input as a sequence of sentiment-enriched trigram vectors, forming a fixed-dimensional matrix that preserves localized contextual and sentiment information. By retaining individual trigram representations rather than collapsing them into a single pooled vector, the framework maintains fine-grained sentiment patterns that are critical for intra-sentential code mixing text. This trigram-based representation provides a compact yet expressive feature space that is particularly well-suited for sentiment analysis of Malay–English code mixing text, where sentiment cues are often localized and influenced by word order and negation effects.

IV. Experiments

This section describes the dataset used; the experiments set up were carried out to evaluate the performance of the proposed framework. The tools and techniques used in model formulation and evaluation are also discussed.

A. Dataset

The dataset for this study is acquired from the MESocSentiment corpus, a publicly available code mixing Malay–English social media dataset introduced by Shamsuddin et al. [50]. The corpus was created to address the scarcity of annotated code mixing sentiment resources in the Malaysian social media context and represents one of the most recent and comprehensive datasets for this task. The original MESocSentiment corpus includes three sentiment categories, which are positive, negative and neutral with a strong class imbalance toward neutral sentiment. To ensure consistency with the proposed feature extraction framework and the sentiment lexicon employed in this study, the dataset is adapted to a binary sentiment classification setting. Only tweets labeled as positive or negative are retained, while neutral instances are excluded. Table I summarizes the class distribution of the MESocSentiment dataset used in this study. The resulting dataset consists of 4,292 code mixing Malay–English tweets with 3,219 positive and 1,073 negative instances.

TABLE I. Dataset

Dataset	Sentiments	
	Positive	Negative
Code Mixing	3,219	1,073

B. Experimental Setup

To provide a comprehensive and fair evaluation of the proposed feature extraction framework, this study compares its performance against two groups of baseline approaches that are widely used in sentiment analysis research: traditional sparse feature representations with linear classifiers and multilingual contextual embedding models with neural classifiers. This design allows the evaluation to cover both classical machine-learning pipelines that remain strong for short text classification and modern transformer-based representations that are commonly applied to multilingual and code mixing sentiment tasks.

First, two traditional feature-based baselines are included to establish strong reference points for sentiment classification on noisy social media text. The first baseline uses a standard bag of words (BoW) style vectorization approach, where code mixing posts are represented using sparse term-based vectors and classified using a linear Support Vector Machine (SVM). The second baseline uses TF-IDF representation with a linear SVM classifier. TF-IDF is particularly suitable for social media sentiment classification because it reduces the influence of frequently occurring conversational fillers and amplifies terms that are more discriminative for class separation. Linear SVM is selected because it has been consistently reported as a competitive classifier for high-dimensional sparse text features and provides a robust and reproducible benchmark for comparison.

Second, two multilingual contextual embedding baselines are included to represent state-of-the-art transformer-based models that capture semantic information beyond surface token frequency. In these baselines, each post is encoded using either mBERT or XLM-R to obtain contextualized representations suitable for downstream sentiment classification. These models are widely adopted in multilingual and cross-lingual NLP tasks and are relevant baselines because code mixing Malay–English text often contains subword variations, spelling inconsistencies and language switching where contextual embedding models are expected to provide more stable representations than sparse features. For classification, a Convolutional Neural Network (CNN) is employed on top of the contextual embeddings for the model to learn discriminative local patterns from the embedding sequences, which is particularly useful for short code mixing social media posts.

For fairness and interpretability, the proposed feature extraction framework uses the same CNN classifier architecture as the contextual embedding baselines. This ensures that performance differences between the proposed approach and transformer-only baselines are primarily attributable to the feature representation strategy rather than differences in classifier capacity. This comparison isolates the contribution of the proposed framework's additional components which are informative trigram filtering, TF-IDF weighting, positional encoding and sentiment polarity scoring when integrated with XLM-R contextual embeddings. As a result, the evaluation framework supports a more direct assessment of whether the proposed feature extraction design provides measurable advantages over both classical sparse feature approaches and standard multilingual embedding pipelines under consistent experimental conditions.

C. Model Parameter for SVM and CNN

For the baseline machine learning models, a linear Support Vector Machine (SVM) classifier was employed due to its

effectiveness and computational efficiency in high-dimensional text feature spaces. The SVM was trained using a linear kernel as implemented in LinearSVC, which is well-suited for sparse representations generated by TF-IDF and CountVectorizer features. The regularization parameter C was fixed at 1.0, providing a balanced trade-off between margin maximization and classification error. No class re-weighting or resampling was applied in order to preserve the natural data distribution. Feature vectors were constructed using word-level unigrams and bigrams, and all SVM parameters were kept constant across folds to ensure fair and reproducible comparisons.

For the deep learning approach, a vanilla Convolutional Neural Network (CNN) architecture was adopted to evaluate the discriminative capability of the proposed and baseline embeddings. The CNN consisted of a single one-dimensional convolutional layer with 128 filters and a kernel size of 3, followed by ReLU activation and global max-pooling to capture the most salient n-gram-level features. A fully connected dense layer with 64 hidden units was used prior to the output layer, and dropout with a rate of 0.5 was applied to mitigate overfitting. The output layer employed a sigmoid activation for binary sentiment classification. The network was optimized using the Adam optimizer with default learning rate settings, and binary cross-entropy was used as the loss function. All CNN models were trained for 10 epochs with a batch size of 16, and identical hyperparameters were maintained across all experimental settings to ensure consistency and comparability. Table II summarizes the model parameter settings for SVM and CNN.

TABLE II. MODEL PARAMETER SETTING FOR SVM AND CNN

Model	Parameter	Setting
Linear SVM	Feature representation	TF-IDF / CountVectorizer
	N-gram range	(1, 2)
	Max features	50,000
	Kernel	Linear
	Regularization (C)	1
	Class weighting	None
	Optimization	Linear hinge loss
	Evaluation	3-fold stratified CV
	Metrics	Accuracy, Precision, Recall, F1-score
CNN	Input embedding	trigram level embeddings
	Embedding dimension	768 / 769
	Convolution filters	128
	Kernel size	3
	Activation	ReLU
	Pooling	Global max pooling
	Dense units	64
	Dropout rate	0.5
	Output activation	Sigmoid
	Loss	Binary cross-entropy
	Optimizer	Adam
	Batch size	16
	Epochs	10
	Evaluation	3-fold stratified CV
	Metrics	Accuracy, Precision, Recall, F1-score

D. Model Performance Evaluation

Performance metrics are used to assess the effectiveness of classification and feature extraction and feature selection techniques. To evaluate the performance of the proposed framework and to compare it with baseline methods, many performance metrics will be implemented depending on the training and testing datasets, including the following evaluation criteria. Four model evaluation metrics were selected: accuracy, precision, recall, and F-measure. The metrics are presented in Eq. (14) to Eq. (17).

Accuracy is the ratio of the correctly classified predictions to the total sum of predictions. It is given as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (14)$$

Precision is the ratio of accurately classified data to the total data classified in the class. It is given as:

$$Precision = \frac{TP}{TP + FP} \quad (15)$$

Recall is the ratio of accurately classified data to the actual data in the class. It is given as:

$$Recall = \frac{TP}{TP + FN} \quad (16)$$

F-measure is the mean of precision and recall. It is given as:

$$F - measure = \frac{2 * Precision + Recall}{Precision + Recall} \quad (17)$$

V. RESULTS AND ANALYSIS

A. Overall Performance

Table III presents the comparative performance of the proposed feature extraction framework against traditional feature-based and multilingual contextual embedding baselines.

TABLE III. PERFORMANCE COMPARISON

Feature Extraction Representation	Classifier	Results			
		Acc	Prec	Rec	F1-score
BoW	SVM	0.879	0.881	0.931	0.924
Tfidf	SVM	0.873	0.871	0.935	0.920
mBERT	CNN	0.862	0.898	0.921	0.909
XLM-R	CNN	0.884	0.905	0.944	0.924
Proposed Framework	CNN	**0.896**	**0.920**	**0.954**	**0.932**

As shown in Table III, among the traditional feature-based approaches, BoW combined with a linear SVM achieves an accuracy of 0.879, a precision of 0.881, a recall of 0.931 and an F1-score of 0.924. The TF-IDF representation with a linear SVM records an accuracy of 0.873, a precision of 0.871, a recall of 0.935 and an F1-score of 0.920. For the multilingual

contextual embedding baselines, the mBERT based CNN model achieves an accuracy of 0.862, a precision of 0.898, a recall of 0.921 and an F1-score of 0.909. The XLM-R based CNN model shows improved performance with an accuracy of 0.884, a precision of 0.905, a recall of 0.944, and an F1-score of 0.924. The proposed feature extraction framework achieves the highest performance across all evaluation metrics. It attains an accuracy of 0.896 and an F1-score of 0.932, outperforming both traditional sparse feature representations and multilingual contextual embedding baselines. The results indicate that the proposed framework provides a more effective representation for sentiment analysis in code mixing Malay–English social media text.

B. Comparison with Traditional Feature-Based Baselines

Traditional feature-based approaches using BoW and TF-IDF combined with a linear SVM classifier demonstrate strong baseline performance, achieving F1-scores of 0.924 and 0.920, respectively. These results confirm that sparse feature representations remain competitive for sentiment analysis on social media text, particularly when sentiment cues are expressed through frequently occurring lexical patterns. However, despite their effectiveness, these approaches rely primarily on surface-level term statistics and do not explicitly model contextual semantics or word order. As a result, they may struggle to distinguish sentiment in code mixing posts where meaning is conveyed through local phrase structure or experiential descriptions rather than isolated keywords. The proposed framework improves upon these baselines by incorporating contextual embeddings and structured feature weighting, resulting in a higher F1-score and improved overall accuracy.

C. Comparison with Multilingual Contextual Embedding Baseline

Among the contextual embedding baselines, the mBERT-based CNN model achieves an F1-score of 0.909, which is lower than both the TF-IDF baseline and the XLM-R-based CNN model. This result suggests that while mBERT provides contextualized representations, it may not sufficiently capture sentiment-relevant patterns in short and informal code mixing social media posts. The XLM-R-based CNN model performs better, achieving an F1-score of 0.924. This improvement reflects XLM-R's stronger multilingual representation capability and its ability to handle cross-lingual lexical variation more effectively than mBERT. Nevertheless, the XLM-R baseline still falls short of the proposed framework, indicating that contextual embeddings alone do not fully capture sentiment discriminative information in code mixing social media text. By integrating additional feature extraction components on top of XLM-R embeddings, the proposed framework achieves further performance gains. This comparison highlights that enriching contextual embedding with sentiment-aware and structure-sensitive features leads to more effective sentiment representation than using transformer embeddings alone.

D. Impact of the Proposed Feature Extraction Components

The performance improvements achieved by the proposed framework can be attributed to the combined effect of its feature extraction components. Informative trigram filtering focuses the representation on sentiment-relevant local contexts that frequently occur in social media complaints or expressions of satisfaction [18,19]. TF-IDF weighting amplifies discriminative sentiment cues while reducing the influence of common conversational fillers that are prevalent in informal online text. Positional encoding preserves short range word order information, which is particularly important in intra-sentential code mixing text, where sentiment may depend on the arrangement of words rather than their presence alone. In addition, sentiment polarity scoring explicitly anchors contextual embeddings toward positive or negative orientation, providing complementary sentiment information that is not always captured by contextual embeddings alone.

The comparison with the XLM-R baseline using the same CNN classifier isolates the contribution of these additional components. Because both models share the same embedding backbone and classifier architecture, the observed performance gains can be directly attributed to the proposed feature extraction strategy.

E. Analysis of Recall and Sensitivity

A notable result in Table III is the high recall value of 0.954 achieved by the proposed framework, which is higher than all baseline models. High recall indicates that the framework correctly identifies a larger proportion of sentiment-bearing instances in the dataset, thereby reducing the number of false negatives. This property is particularly relevant for sentiment analysis of social media text, where users frequently describe situations or experiences instead of explicitly stating emotional reactions. In such cases, sentiment is conveyed through contextual and structural cues rather than isolated polarity words.

The proposed framework demonstrates increased sensitivity to such sentiment-bearing instances by retaining sentiment-relevant trigrams, emphasizing discriminative terms through TF-IDF weighting and incorporating sentiment polarity scoring alongside contextual embeddings. To illustrate this behavior, consider the code mixing Malay–English sentence *"lagi stress liao"* from the dataset. Although the sentence is short and does not contain explicit sentiment adjectives, it conveys dissatisfaction through the combination of the English term *"stress"* with Malay and colloquial discourse markers. At the trigram level, this local context is preserved while the contextual representation generated by XLM-R captures the negative experiential state associated with stress, enabling the classifier to correctly predict negative sentiment.

F. Error Analysis

Despite its performance, the proposed framework's result shows an error rate of approximately 10.4%, indicating that a small portion of code mixing social media posts remains misclassified. Analysis of these errors reveals several recurring challenges. First, some posts contain mixed sentiment cues where positive and negative signals appear within the same short sentence. In such cases, competing cues may influence the aggregated trigram representation and sentiment polarity score, leading to ambiguous predictions. Second, the informal nature of social media text introduces slang, abbreviations, and filler expressions that do not carry sentiment but affect local context representation. Short or fragmentary sentence structures further reduce the amount of contextual information available for

reliable classification. Finally, pragmatic language use, including sarcasm and irony, remains a challenge. In some instances, positive lexical items are used in a critical or sarcastic manner, which may not be fully captured by sentiment polarity scoring or contextual embeddings.

Despite strong overall performance, some errors remain as mentioned. For instance, the sentence "boleh login akhirnya tapi ambil masa lama" contains both a positive outcome ("boleh login akhirnya") and a negative experience ("ambil masa lama"). The presence of competing sentiment cues within a short sentence may lead to ambiguous representations and misclassification. Other example, sentences such as "ok boring gila lah" and "jom tengok kes otosan gaming biar tak padu asalkan" contain conversational fillers ("ok", "jom", "lah") and social expressions that contribute limited sentiment information. The dominance of such informal tokens reduces the amount of sentiment-relevant context available to the model, making classification more sensitive to a small number of words or trigrams. Similarly, sarcastic expressions such as "good you" in sentences like "intercontinental hotel digital perak got hacked, good you" can be misinterpreted due to the positive lexical form being used with negative intent. These limitations are consistent with challenges reported in prior work on sentiment analysis of social media text and highlight areas for future improvement.

VI. Conclusion

Overall, the results demonstrate that the proposed feature extraction framework provides consistent and meaningful improvements over both traditional sparse feature representations and multilingual contextual embedding baselines on code mixing Malay–English social media text. By combining contextual embeddings with trigram-level context, statistical weighting, positional structure and sentiment polarity score, the framework achieves more robust sentiment representation and improved classification performance in a realistic social media setting. Despite these improvements, several limitations remain. The framework relies on a predefined sentiment lexicon, which may not fully capture evolving slang and domain-specific expressions commonly found in social media. In addition, sarcasm and implicit sentiment expressions remain challenging due to their contextual and pragmatic nature. Future work will focus on extending the framework to incorporate dynamic lexicon expansion, sarcasm-aware modeling and cross-domain generalization. The evaluation of the framework will also be extended to cross-domain and multi-dataset evaluation to validate the proposed framework's generalization capability.

Acknowledgment

The authors would like to thank the Public Service Department (PSD), the Ministry of Higher Education (MOHE) and the Universiti Teknologi Malaysia (UTM) for their educational and financial support. This work is conducted at the Faculty of Artificial Intelligence (FAI).

References

[1] M. Ali, K. S. Teja, N. Gupta, P. Patwa, A. Chatterjee, V. Jain, and A. Das, "CONFLATOR: Incorporating Switching Point based Rotatory Positional Encodings for Code-Mixed Language Modeling," arXiv:2309.05270, 2023.

[2] M. N. Sadiku and C. M. Akujuobi, "The Internet," in Fundamentals of Computer Networks. Cham, Switzerland: Springer, 2022, pp. 51–69.

[3] K. Ibragimov, "Internet's potential as a source of legal information," Science and Innovation, vol. 2, no. C8, pp. 36–39, 2023.

[4] Z. E. Zamri, J. Yusof, Y. F. Yasin, A. Ibrahim, S. A. Panatik, H. A. Kean, and N. A. K. Zamri, "Smartphone addiction and loneliness among students," Jurnal Kemanusiaan, pp. 85–94, 2023.

[5] D. Guo, H. Chen, R. Wu, and Y. Wang, "AIGC challenges and opportunities related to public safety: A case study of ChatGPT," Journal of Safety Science and Resilience, 2023.

[6] T. A. Al-Qablan, M. H. Mohd Noor, M. A. Al-Betar, and A. T. Khader, "A survey on sentiment analysis and its applications," Neural Computing and Applications, vol. 35, no. 29, pp. 21567–21601, 2023.

[7] N. H. Mahadzir, N. H. A. Razak, and M. F. M. Omar, "A New Sentiment Analysis Model for Mixed Language using Contextual Lexicon," in Proc. IEEE Int. Conf. Recent Advances and Innovations in Engineering (ICRAIE), 2020, pp. 1–5.

[8] C. Tho, Y. Heryadi, L. Lukas, and A. Wibowo, "Code-mixed sentiment analysis of Indonesian and Javanese language using lexicon-based approach," Journal of Physics: Conf. Series, vol. 1869, no. 1, p. 012084, 2021.

[9] J. Mountstephens, M. T. Z. Quen, and L. Hung, "Bilingual sentiment analysis on Malaysian social media using VADER and normalisation heuristics," Journal of Theoretical and Applied Information Technology, vol. 101, no. 12, 2023.

[10] J. T. Kong, F. H. Juwono, I. Y. Ngu, I. G. D. Nugraha, Y. Maraden, and W. K. Wong, "A Mixed Malay–English Language COVID-19 Twitter Dataset: A Sentiment Analysis," Big Data and Cognitive Computing, vol. 7, no. 2, p. 61, 2023.

[11] N. P. Romadhona, S. E. Lu, B. H. Lu, and R. T. H. Tsai, "BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset," in Proc. Int. Conf. Computational Linguistics, 2022, pp. 4418–4428.

[12] P. Kodali, V. Shivkumar, S. Joshi, M. Choudhary, P. Kumaraguru, and M. Shrivastava, "Adapting Multilingual Models to Code-Mixed Tasks via Model Merging," arXiv:2510.19782, 2025.

[13] L. Zeng, "Leveraging large language models for code-mixed data augmentation in sentiment analysis," arXiv:2411.00691, 2024.

[14] A. Kumar, A. Pandey, S. Ahlawat, and Y. Prasad, "On Enhancing code-mixed sentiment and emotion classification using FNet and FastFormer," arXiv, 2024.

[15] Y. Feng, F. Li, and P. Koehn, "Toward the limitation of code-switching in cross-lingual transfer," in Proc. EMNLP, 2022, pp. 5966–5971.

[16] E. Hashmi, S. Y. Yayilgan, and S. Shaikh, "Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers," Social Network Analysis and Mining, vol. 14, no. 1, p. 86, 2024.

[17] M. Krasitskii, O. Kolesnikova, L. C. Hernandez, G. Sidorov, and A. Gelbukh, "Advancing sentiment analysis in Tamil-English code-mixed texts," arXiv:2503.23295, 2025.

[18] E. Kasmuri and H. Basiron, "Subjectivity analysis of an enhanced feature set for code-switching text," IJACSA, vol. 15, no. 9, 2024.

[19] A. Perera and A. Caldera, "Sentiment analysis of code-mixed text: A comprehensive review," 2024.

[20] P. A. Joshi, V. M. Pathak, and M. R. Joshi, "Sentiment analysis from social media data in code-mixed languages using machine learning classifiers," in Proc. Int. Conf. Data Science, 2023, pp. 203–222.

[21] K. Shanmugavadivel, S. H. Sampath, P. Nandhakumar, P. Mahalingam, M. Subramanian, P. K. Kumaresan, and R. Priyadharshini, "An analysis of machine learning models for sentiment analysis of Tamil code-mixed data," Computer Speech & Language, vol. 76, p. 101407, 2022.

[22] P. Ghosh, S. Vashishth, R. Dabre, and P. Bhattacharyya, "A morphology-based investigation of positional encodings," arXiv:2404.04530, 2024.

[23] J. Kim, Y. Na, K. Kim, S. R. Lee, and D. K. Chae, "SentiCSE: A sentiment-aware contrastive sentence embedding framework," in Proc. LREC-COLING, 2024, pp. 14693–14704.

[24] R. Petrolito and F. Dell'Orletta, "Word embeddings in sentiment analysis," Computational Linguistics CLiC-it, 2018.

[25] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava, "Sentiment analysis of code-mixed languages leveraging resource rich languages," in Proc. CICLing, 2018, pp. 104–114.

[26] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava, and P. Koehn, "De-mixing sentiment from code-mixed text," in Proc. ACL Student Research Workshop, 2019, pp. 371–377.

[27] K. Yadav, A. Lamba, D. Gupta, A. Gupta, P. Karmakar, and S. Saini, "Bi-LSTM and ensemble-based bilingual sentiment analysis," in Proc. IEEE INDICON, 2020, pp. 1–6.

[28] Y. K. Wiciaputra, J. C. Young, and A. Rusli, "Bilingual text classification using XLM-R," Int. J. Advances in Soft Computing, vol. 13, no. 3, 2021.

[29] C. Kumaresan and P. Thangaraju, "ELSA: Ensemble learning based sentiment analysis," Measurement: Sensors, vol. 25, p. 100663, 2023.

[30] T. Tang, X. Tang, and T. Yuan, "Fine-tuning BERT for multi-label sentiment analysis," IEEE Access, vol. 8, pp. 193248–193256, 2020.

[31] N. Sabri, A. Edalat, and B. Bahrak, "Sentiment analysis of Persian-English code-mixed texts," in Proc. IEEE CSICC, 2021, pp. 1–4.

[32] Z. Jin, X. Lai, and J. Cao, "Multi-label sentiment analysis based on BERT with modified TF-IDF," in Proc. IEEE ISPCE-CN, 2020, pp. 1–6.

[33] J. Mutinda, W. Mwangi, and G. Okeyo, "Lexicon-enhanced BERT embedding with CNN," Applied Sciences, vol. 13, no. 3, p. 1445, 2023.

[34] A. Patil, V. Patwardhan, A. Phaltankar, G. Takawane, and R. Joshi, "Comparative study of pre-trained BERT models for code-mixed data," in Proc. IEEE I2CT, 2023, pp. 1–7.

[35] G. Takawane, A. Phaltankar, V. Patwardhan, A. Patil, R. Joshi, and M. S. Takalikar, "Leveraging language identification to enhance code-mixed text classification," arXiv:2306.04964, 2023.

[36] V. Patwardhan, G. Takawane, N. Kelkar, O. Gaikwad, R. Saraf, and S. Sonawane, "Sentiment analysis of Marathi-English code-mixed data," in Proc. IEEE ESCI, 2023, pp. 1–5.

[37] C. M. Liapis, A. Karanikola, and S. Kotsiantis, "Enhancing sentiment analysis with distributional emotion embeddings," Neurocomputing, vol. 634, p. 129822, 2025.

[38] E. Kasmuri and H. Basiron, "Building a Malay-English code-switching subjectivity corpus for sentiment analysis," Int. J. Advance Soft Computing Applications, vol. 11, no. 1, 2019.

[39] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Z. Abidin, "Developing cross-lingual sentiment analysis of Malay Twitter data," IJACSA, vol. 10, no. 1, 2019.

[40] M. J. Fuady and R. Ibrahim, "Multilingual sentiment analysis on social media disaster data," in Proc. IEEE ICEEIE, 2019, pp. 269–272.

[41] A. He and M. Abisado, "Text sentiment analysis using BERT-CNN-BiLSTM-Att model," IEEE Access, vol. 12, pp. 45229–45237, 2024.

[42] M. S. M. Suhaimin, M. H. A. Hijazi, and E. G. Moung, "MSSThreD: Multitask social media sentiment analysis with sarcasm detection," in Proc. IEEE ICITACEE, 2024, pp. 25–30.

[43] [43] Y. F. Tan, H. S. Lam, A. Azlan, and W. K. Soo, "Sentiment analysis for Telco popularity on Twitter," in ICADIWT, pp. 112–125, 2016.

[44] D. H. Wahid and S. N. Azhari, "Peringkasan sentimen ekstraktif di Twitter menggunakan hybrid TF-IDF," IJCCS, vol. 10, no. 2, pp. 207–218, 2016.

[45] Z. Husein, "Malay Dataset," GitHub repository. [Online]. Available: https://github.com/huseinzol05/Malay-Dataset. Accessed: 2026.

[46] K. S. M. Anbananthen, S. Selvaraju, and J. K. Krishnan, "The generation of Malay lexicon," American Journal of Applied Sciences, vol. 14, pp. 503–510, 2017.

[47] K. Hämmerl, J. Libovický, and A. Fraser, "Combining static and contextualised multilingual embeddings," arXiv:2203.09326, 2022.

[48] A. Fernando and S. Ranathunga, "Linguistic entity masking for low-resource languages," Knowledge and Information Systems, vol. 67, no. 11, pp. 9905–9946, 2025.

[49] I. Puranegedara, T. Chathumina, N. Ranathunga, N. De Silva, S. Ranathunga, and M. Thayaparan, "Utilizing multilingual encoders to improve LLMs," in Proc. IEEE MERCon, 2025, pp. 641–646.

[50] A. M. Shamsuddin, S. S. Juan, S. Chua, and A. Bramantoro, "Semi-Automatic Sentiment Identification for Malay-English Code-Switched Data," Journal of Advanced Research Design, vol. 123, no. 1, pp. 198–212, 2024.