Sentiment Analysis in Code Mixing Malay Text: A Review

Surendran Selvaraju
Razak Faculty of Technology & Informatics
University Technology Malaysia
Malaysia
surendran90@graduate.utm.my

Nilam Nur Amir Sjarif
Razak Faculty of Technology & Informatics
University Technology Malaysia
Malaysia
nilamnur@utm.my

Mohd Syahid Mohd Anuar
Razak Faculty of Technology & Informatics
University Technology Malaysia
Malaysia
syahid.anuar@utm.my

2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS) | 979-8-3315-2855-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIDAS63860.2024.10730298

Abstract— **Sentiment analysis in code mixing received a significant amount of attention in recent years. Code mixing involved two or more languages in a text. Significant gaps in existing sentiment analysis models conventionally cater to monolingual only. These conventional models often fail to capture the nuanced expressions found in code mixing text. Sentiment analysis of Malay text has drawn a lot of attention from scholar especially in online platforms. The syntax, grammar and semantic rules of Malay are varying than English. Additionally, code mixing text also includes slang, abbreviation and non-standard language variations. This adds to the complexities in understanding the context and meaning of the code mixing text. Therefore, this review paper investigated and presents sentiment analysis in code mixing briefly. Different approaches are also discussed in this paper from previous works, ranging from lexicon based approach to advanced machine learning based approach. Lastly, this review highlights the unique challenges and significant opportunities in sentiment analysis for code mixing Malay text.**

Keywords—Code Mixing Malay Online Text, Sentiment Analysis, Preprocessing, Feature Extraction, Feature Selection, Sentiment Analysis

I. Introduction

The practice of mixing different varieties of Malay language is also becoming increasingly common and this is true especially on online platforms. People continue to express their opinions and participate in political, sports and entertainment discussions on a variety of platforms both online and offline. This is due to the topics receiving a lot of attention from the public [1][2][3]. As a result, code mixing Malay text is becoming more common. In code mixing Malay text, words and phrases from various ethnic languages such as English, Mandarin, and Tamil are seamlessly integrated into the Malay language. The presence of code mixing text on online platforms primarily have highlights the importance of developing efficient techniques for sentiment analysis in order to understand the emotion and sentiment that are being expressed in a such text. It has received a significant amount of attention in recent years as a result of the growing availability of user-generated content on online platforms and the need to extract valuable insights from the vast amounts of textual data [4]. It can be used to analyse the generated text data to uncover hidden insights. Sentiment analysis is commonly used by researchers and businesses to analyse customer feedback, survey responses and public reviews on product and service. While researchers have paid close attention to sentiment analysis on Malay text, there has been little exploration of sentiment analysis techniques specifically tailored for code mixing Malay text [5][6][8].

This gap is evident as most studies have focused on either monolingual sentiment analysis or general multilingual approaches without addressing the unique challenges of code mixing. For instance, Romadhona in [5] emphasize the need for a dedicated code mixing Malay corpus which highlighting the scarcity of resources tailored for this context. Similarly, Mountstephens [6] and Kong [7] underscore the complexities of handling code mixing text and pointing out the limitations of current methodologies.

Existing sentiment analysis models are often designed for single languages thus it requires adaptation or development of specific techniques for code mixing text. Code mixing introduces complexity as it involves two or more languages in a text. The syntax, grammar, and semantic rules can vary significantly between Malay and the language it is mixed with. This leads to complexities in understanding the context and meaning of the text. There is a lack of comprehensive datasets and resources specifically designed for code mixing Malay text [4][5][6]. In addition, code mixing text also contains slang, abbreviations and non-standard language variations. This informal text can be especially difficult to analyse because they may have different sentiment expressions and may not be well-represented in standard datasets. Regardless of the fact that sentiment analysis has been extensively researched in a variety of languages, the unique characteristics of code mixing Malay text present both challenges and opportunities for researchers.

The following section will discuss on what is sentiment analysis in Section II and various approach of previous works for code mixing text and code mixing Malay text in Section III and IV respectively. Section V presented on how data collection activity is carried out for code mixing Malay text while Section VI discussed on the performance of each approach used in code mixing Maly text by previous researchers. And the last section VII conclude this review paper.

II. Sentiment Analysis

Sentiment refers to an individual's subjective response, personal encounter, emotional state or viewpoint [1]. Sentiment analysis is the process of detecting and monitoring the opinions articulated by individuals on the internet. For code mixing text, the process begins with preprocessing steps like noise removal, tokenization, and language identification.

To address data scarcity, data augmentation techniques are usually employed. In lexicon-based approach, each word or phrase in the text is matched against the sentiment lexicon [11][12]. The sentiment scores of the matched words are aggregated to determine the overall sentiment of the text. In advanced machine learning-based approach, the text is transformed into numerical vectors using feature extraction methods such as Bag of Words, TF-IDF, word embeddings, and contextual embeddings like BERT [14][15]. These vectors are used to train machine learning models enabling them to recognize patterns and nuances in the mixed language data.

III. Sentiment analysis

In this review, this research examines recent study on approaches used for code mixing text. Most of the researches on code mixing text concentrated on the English-Hindi context [8][9][10]. It is observed that recent researches on code mixing text have opted to advanced machine learning based approach.

Choudhary et al. [8] proposes a novel approach called Sentiment Analysis of Code-Mixed Text (SACMT) to classify sentences into their corresponding sentiment of positive, negative or neutral using contrastive learning. The author applies the shared parameters of siamese networks to map the sentences of code mixing and standard languages to a common sentiment vector. The researcher used skip-gram vectors to generate word embedding. The sentiment analysis task is demonstrated using a combination of word embeddings and convolutional neural networks (CNNs) with Siamese Bi-LSTM and a fully connected layer. Lal et al. [10] proposed a hybrid model for sentiment analysis (SA) tasks in English-Hindi code mixing text using sub-word embedding. The authors began by employing a Convolutional Neural Network (CNN) architecture to generate sub-word level representations for the sentences. These representations are then used as inputs to a Dual Encoder Network which is comprised of two distinct Bi-LSTM. This approach combined feature network that contained orthographic features in addition to word embedding that had been specifically trained.

Similar Bi-LSTM model is utilised by Yadav et al. [12] used sequence features on code mixing text in English-Hindi. The code mixing data is pre-processed and tokenized. The Bi-LSTM model's training is done in two different sequences. The first sequence is the input pattern and the second sequence is a mirror image of the input pattern. This approach has the potential to continue providing the network with additional context which can lead to learning the actual problem more quickly and in detail. The resulting word embedding is inputted to a classifier for sentiment analysis task. Due to the model's capacity to capture sequential information, it is observed that the model performs better for sentences that are longer in length. Slightly, different approach has been opted by the author in [13] using Bi-LSTM for carrying out sentiment analysis on Dravidian code mixing text. The authors converted the text data into feature vector and then fed it into a Bi-LSTM network. Word2vec and skip-gram model is used to generate the feature vector also known as word embedding.

Yadav and Chakraborty in [14] proposed a zero-shot approach to solve sentiment analysis task on code mixing Spanish-English text. They used multilingual and crosslingual embeddings to transfer knowledge from monolingual text to code mixing text. The authors emphasized that a significant amount of work is required in bilingual, crosslingual and multilingual feature extraction for text to facilitate model building for low resource languages. In this approach, they first independently learn monolingual word representations of L1(English) and L2 (Spanish) from large monolingual corpus and then learn a transformation matrix W to map vectors from one language to the vector of the other language. The monolingual embeddings for English and Spanish used in these experiments were obtained from Grave et al. [15]. Both supervised and unsupervised variants are available through Modularizing Unsupervised Sense Embeddings (MUSE) [16]. Later, the crosslingual fastText embedding is used as input for the sentiment analysis task performed by the Bi-LSTM network. According to the findings of the research, training on an actual code mixed corpus and making use of subwords are both helpful in the process of developing better embeddings. This is due to the fact that simply concatenating monolingual or parallel corpora is not enough to capture the syntactic and semantic features of code mixing text. Kusampudi et al. [17] proposed unsupervised data normalization for sentiment analysis in code mixing Telugu-English text. The normalization techniques involved are normalized transliteration variants and spelling errors. Later, the normalized text is used to generate word vectors. N-grams and Term Frequency and Inverse Document Frequency (TF-IDF) are used to create the word vectors. These vectors are then passed to multilayer perceptron (MLP) model for sentiment analysis task.

In their study, [18] state that different sentiments may exist together and the proportion of each sentiment in the code mixing text is often unbalanced. Inspired by the recently proposed BERT model the author investigates how to fine-tune BERT for multi-label sentiment analysis for unbalanced code mixing Chinese-English text. Because Chinese-English text is classified as under resource text, the authors employ multi-label BERT for this task and conduct experiments with the pre-trained model and parameters. To obtain balanced samples, they use data augmentation technique based on translation and a multiple undersampling method by taking into account the characteristics of the unbalanced dataset and the scarcity of text data availability. These samples are trained by several independent multi-label BERT classifiers and the outputs of these classifiers are combined using ensemble learning. Kannan et al. [19] proposes an approach that utilizes IndicBERT another variant of BERT. To attain better performance of the BERT model, the model parameters are fine-tuned with code mixing text to obtained optimum learning rate.

Sabri in [20] works on the collection, labelling and subsequent creation of a dataset consisting of tweets that mix Persian and English language. mBERT is used to extract word embedding for the code mixing text. To automatically learn the polarity scores of these tweets, an ensemble model that is comprised of three Bi-LSTM networks is proposed. Additionally, it makes use of mBERT pretrained embeddings and translation models. The Bi-LSTM network has an attention mechanism added to it which enables the model to focus more on the key words in the sentence and pay less attention to the less significant words.

Wiciaputra, Young, and Rusli in [21] proposes a method for bilingual text classification in English and Indonesian using XLM-RoBERTa transfer learning. The research addresses the problem of classifying text data in multiple

languages with the goal of achieving high accuracy in both English and Indonesian classification tasks. The authors fine-tune XLM-RoBERTa model on a large-scale bilingual dataset of English and Indonesian text with various classification categories. According to the findings of the study, transfer learning with XLM-RoBERTa is an effective method for bilingual text classification outperforming other models in terms of accuracy and classification performance. The results demonstrate the effectiveness of using pretrained multilingual model for cross-lingual text classification tasks.

[22] works on pretrained BERT model and intended to perform sentiment analysis for low resource code mixing text of Tamil and English. The authors experimented on Adapter-BERT model [27]. Adapter-BERT is fine-tuned by inserting a two-layer fully connected network which is known as an adapter. This adapter is inserted into each transformer layer of the BERT model. During the end-task training, only the adapters and the connected layer are trained while no other BERT parameters are altered. In [23], the author also aims to investigate the performance of different pre-trained BERT models specifically in the context of code mixing Hindi-English text. They compare various variants of BERT such as HingBERT, RoBERTa, AlBERT, and mBERT to determine the model's effectiveness in handling code mixing text in sentiment analysis. The study compared different models on various datasets and reported state-of-the-art performance using HingBERT-based model. The research highlights the importance of code mixing language model and the need for more data and research in code mixing Hindi-English text.

To extend the previous researcher's approach, [24] perform sentiment analysis on code mixing Hindi-English text using HingBERT and other BERT variants by proposing an approach to improve the performance of BERT-based models by leveraging language tagging features. Interleaved word level language tagging and adjacent sentence level language tagging are the two methods of language expansion that are investigated by the authors. They demonstrate that language tagging augmentation improves the performance of all BERT variants by evaluating the effectiveness of various BERT models on the code mixing text and finding that the improved performance is due to language augmentation. The results highlight the effectiveness of the proposed approach across BERT models. Patwardhan et al. [25] used the same approach to study sentiment analysis of code mixing on Marathi-English text. The researchers include POS tagging in the text annotation process in addition to language tagging. Each word in the text will be tagged with the appropriate language and POS tag. Later, this annotated text will be fed into SVM and Naive Bayes machine learning models to perform sentiment analysis and their performances are evaluated. The results showed that both models performed well with the SVM model performing slightly better. For future improvement, the author recommends using advanced machine learning based approach such as BERT and RoBERTa for sentiment analysis on low resources language.

A novel method referred to as ELSA (Embedding Layer Sentiment Analysis) is proposed by the author [26] for the purpose of conducting sentiment analysis on diverse text, more specifically tweets that contain code mixing of Tamil-English text. This method classifies tweets into positive, negative and neutral categories by combining Generative Adversarial Networks (GAN) and Self-Attention Networks (SAN). For the purpose of conducting multilingual sentiment analysis, the authors suggest employing a word embedding that is based on XLM-R. The word embedding is then later fed to the ensemble classifier (ELSA). XLM-R excels at understanding distinct languages separately but code mixing text blends languages within sentences and phrases. The author emphasizes that this feature of code mixing text presents unique challenges.

IV. SENTIMENT ANALYSIS ON MALAY CODE MIXING TEXT

Code mixing Malay text sentiment analysis is an emerging field of research that aims to analyse the sentiment of mixed Malay language text. Code mixing in Malay language is still an understudy language. Study in this language can be categorized to lexicon based and advanced machine learning based approaches.

Zabha et al. [28] present a study that employed a lexicon based approach to develop a cross-lingual sentiment analysis in order to address the scarcity of annotated corpora. The proposed method is lexicon based with sentiment lexicons constructed in both Malay and English. In both languages, the researchers manually compile sentiment words and their corresponding sentiment polarities. These lexicons are then applied to the words and phrases found in the code mixing Twitter text to assign sentiment scores. For feature selection, the study employs frequent noun or noun phrase identification and association mining with heuristic-guided pruning. The limitations of the lexicon based approach are discussed, including the difficulty of dealing with slang, abbreviated words and dialect as well as the need for continuous updates to the sentiment lexicons for effective words identification.

Fuady and Ibrahim in [29] conduct sentiment analysis on code mixing Malay text from social media disaster data. The proposed sentiment model is divided into three phases which are data preprocessing, learning and classification. The learning phase includes data preprocessing and machine translation resulting in a comparable multilingual corpus. The multilingual corpus is built using the English and Malay Wikipedia sites. The authors used word2Vec and Ballahurs' approaches to learn a multilingual word embedding using the multilingual corpus. The authors then used a multilayer deep learning approach to train the model and used the multilingual word embedding in the embedding layer. The shortcoming of this approach is its ineffectiveness in handling unknown words due to morphological similarity during the learning phase on word embedding.

Mahadzir et al. [30] proposed a contextual lexicon approach to address ambiguity by providing contextual information in classifying the sentiments of code mixing Malay text. This is in contrast to previous research which found that the majority of researchers used translation to a single language. The WordNet database is used to extract word senses from WordNet glosses based on context. The model analyses the sentiment expressed in the mixed language text using the contextual lexicon. Different approach applied by [31] by translating Indonesian-Javanese code mixing text to standard language English. The authors are also determined to address the challenges of sentiment analysis in code mixing text, including language variation and the lack of annotated data. Later the translated text polarity is extracted using SentiNetWord and Vader. The sentiment of the text is calculated using a simple mathematical formula. Both the authors realized that translating code mixing text is

challenging due to the complexity of language variation and the lack of orthographic information. Ahmad & Abdullah in [32] taking a new approach using hybrid corpus-based approach and support vector machine (SVM) to analyse sentiment on Malaysia budget comments on social media. The sentiment corpus based combines Malay and English words together. Sentiment corpus based is chosen because it gives the sentiment values directly and suitable for domain specific data. Later, the sentiment corpus is used to train the model to identify the sentiment of each tweet related to Malaysia Budget. The author further mentions that the SVM parameters need to be further explored and more complex model for analysis can be applied to improve the accuracy of sentiment analysis on code mixing Malay text.

Based on previous research, Romadhona et al. [5] working with the same objective to build code mixing corpus. This code mixing corpus is aiming to establish a model that is able to deal with code mixing Malay text. The author uses data augmentation technique to construct the first ever code mixing Malay corpus. Sentences in the corpus uses three languages which are English, Malay, and Chinese. Later, this corpus was used to train XLM language model from scratch and this newly trained model is known as Mixed XLM. The main difference between this proposed Mixed XLM and vanilla XLM is it is able to automatically recognize the language of each input token and handles code mixing input. The author develops a language tagging algorithm to automatically label the language of each word in the input.

Mountstephens et al. [6] addresses code mixing Malay issue by using a lexicon based approach. The author focusses on sentiment analysis in Malaysian social media by addressing bilingual and code mixing challenges in Malay and English. The approach involves translating the English VADER lexicon to Malay using both automatically and manually to build a code mixing sentiment analysis system. For feature extraction, it uses VADER's lexicon including sentiment-laden words and emoticons adjusted for Malaysian context. The work involves identifying key sentiment words in both languages and implementing normalization heuristics for text peculiarities in Malaysian social media. The study evaluates the performance of this bilingual system against other machine learning algorithms and the standalone Malay VADER system. The work contribution lies in its novel application of the VADER lexicon to code mixing setting and the implementation of normalization heuristic to cater to social media in Malaysian context. However, the heuristics developed are static and may not be efficient in capturing complex code mixing patterns.

Kong et al. in [8] focusing on methods to solve the identification of rare and unknown words in code mixing Malay text. For this purpose, a data compression method using Byte-Pair Encoding (BPE) is applied on the tweet texts and two deep learning approaches were used. BPE tokenization is used to encode rare and unknown words into smaller meaningful subwords. The author introduces novel modality, the labelled tweets are converted into image files and CNN is used for the sentiment analysis task. This is the first ever approach using image for sentiment analysis on Malay text. Experiments were conducted to explore different BPE vocabulary sizes with the BPE-Text-to-Image-CNN and BPE-M-BERT models. Overall, the results show that BPE-M-BERT slightly outperforms the CNN model, thereby showing that the pre-trained mBERT model has the advantage for processing multilingual texts.

Contextual word embedding can be used for code mixing sentiment analysis with no fine-tuning and less resources, improving accuracy and reducing training time [5][6][7]. Therefore, this research proposed sentiment analysis scheme based on BERT along with other feature extraction and feature selection techniques to obtain better performance than most of the models of previous research. The following section will discuss the proposed scheme for code mixing text.

V. DATASET COLLECTION

In the data collection phase for sentiment analysis of code mixing Malay text, a robust and diverse dataset is essential to ensure comprehensive coverage of language usage and sentiment expression. The data is sourced from various social media platforms, online forums and communication channels where code mixing is prevalent. Due to scarcity in available annotated data source for this language, the existing annotated monolingual Malay data is augmented to create annotated code mixing Malay text [5][29][30].

VI. DISCUSSION

Challenges and significant opportunities in sentiment analysis for code mixing Malay text Code mixing in Malay text refers to the practice of mixing Malay with other languages in written or spoken communication. Research in code mixing Malay text has identified several challenges and research gaps in this area. One of the primary challenges is the lack of comprehensive linguistic resources and tools specific to code mixing Malay text. Examples of such resources are sentiment lexicons and language model trained on code mixing data.

Initially, research in code mixing Malay text often relied on lexicon based methods for sentiment analysis [1][28][30]. This involved using sentiment lexicons or dictionaries to identify sentiment bearing words and calculate sentiment scores. However, these methods were limited in capturing the nuanced sentiment expressed in code mixing text as they are unable to consider the contextual information and linguistic variations inherent in code mixing. Although a contextual lexicon approach been proposed by [30] but this approach heavily relies on the availability and quality of the lexicon. If the lexicon does not contain sentiment information for all the words and phrases in the code mixing text, the approach may not accurately capture the sentiment expressed in the text. Additionally, the lexicon approach may not be effective in handling new or emerging words and phrases that are not present in the lexicon. This limitation can be addressed by continuously updating and expanding the lexicon to include new words and phrases. However, this requires significant effort and resources.

As research progressed, there was a shift towards utilizing static word embeddings such as Word2Vec to capture semantic relationships between words in code mixing Malay text. These embeddings provided a more nuanced representation of words and allowed for capturing some level of context. However, they still struggled to fully capture the complex linguistic patterns and context-specific sentiments in code mixing Malay text. As mentioned by Fuady and Ibrahim in [29], the shortcoming of this approach is its ineffective in handling unknown words due to morphological similarity.

More recently, research in code mixing Malay text has embraced the use of advanced machine learning approach, particularly language models such as BERT (Bidirectional Encoder Representations from Transformers) and its variants M-BERT [7] and XLM [5]. These models generate contextualized word embeddings that encode the meaning of words based on their surrounding context. Contextualized word embeddings have shown significant advancements in handling code mixing text by capturing complex linguistic patterns and context-specific sentiment.

However, there is still a gap in the research on using contextual word embeddings of BERT for sentiment analysis in code mixing Malay text. There is little research on using contextual word embedding of BERT for sentiment analysis in code mixing Malay text. Sentiment analysis of code mixing text can be improved by leveraging feature extraction and feature selection techniques alongside BERT [11][13]. Utilizing this word embedding generated from BERT can helps to further research for the under resource languages such as Malay-English [5][7]. Contextual word embedding can be used for code mixing sentiment analysis with require no fine-tuning and less resources, improving accuracy and reducing training time [11][13][14].

VII. CONCLUSION

In conclusion, this review highlights the unique challenges and significant opportunities in sentiment analysis for code mixing Malay text. While existing sentiment analysis models often cater for standard language, they face significant challenges when confronted with the linguistic complexities of code mixing. Hence, different approaches discussed in this paper from previous works ranging from lexicon based approach to advanced machine learning approach leveraging language models illustrate a promising edge for improving sentiment analysis in multilingual contexts. As we continue to bridge the gap in resources and refine models adapted for code mixing text, the possibility of improving understanding of nuanced linguistic expressions in Malay-English interactions grows. This has significant implications for both academic research and practical applications in multilingual society.

ACKNOWLEDGEMENT

The authors extend their appreciation to the Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia (UTM) for their educational and financial support. This work is conducted at the Faculty of Artificial Intelligence. This research is financially supported by Universiti Teknologi Malaysia under UTMER Grant Vot Number Q.K130000.3856.31J90.

REFERENCE

[1] Kasmuri, E., & Basiron, H. (2019). Building a Malay-English code-switching subjectivity corpus for sentiment analysis. *Int. J. Advance Soft Compu. Appl*, *11*(1).

[2] Thara, S., & Poornachandran, P. (2022). Social media text analytics of Malayalam–English code-mixed using deep learning. *Journal of big Data*, *9*(1), 45.

[3] Choudhary, H., Shukla, M., & Raghavendra, S. (2023). Predicting bitcoin price fluctuation by Twitter sentiment analysis. In *Recent Trends in Computational Sciences* (pp. 77-83). CRC Press.

[4] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., ... & Yang, M. H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, *56*(4), 1-39.

[5] Romadhona, N. P., Lu, S. E., Lu, B. H., & Tsai, R. T. H. (2022, October). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4418-4428).

[6] MOUNTSTEPHENS, J., QUEN, M. T. Z., & HUNG, L. (2023). BILINGUAL SENTIMENT ANALYSIS ON MALAYSIAN SOCIAL MEDIA USING VADER AND NORMALISATION HEURISTICS. *Journal of Theoretical and Applied Information Technology*, *101*(12).

[7] Kong, J. T., Juwono, F. H., Ngu, I. Y., Nugraha, I. G. D., Maraden, Y., & Wong, W. K. (2023). A Mixed Malay–English Language COVID-19 Twitter Dataset: A Sentiment Analysis. *Big Data and Cognitive Computing, 7*(2), 61.

[8] Choudhary, N., Singh, R., Bindlish, I., & Shrivastava, M. (2018, March). Sentiment analysis of code-mixed languages leveraging resource rich languages. In *International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 104-114). Cham: Springer Nature Switzerland.

[9] Mukherjee, S. (2019, December). Deep learning technique for sentiment analysis of hindi-english code-mixed text using late fusion of character and word features. In 2019 IEEE 16th India Council International Conference (INDICON) (pp. 1-4). IEEE.

[10] Lal, Y. K., Kumar, V., Dhar, M., Shrivastava, M., & Koehn, P. (2019, July). De-mixing sentiment from code-mixed text. In Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop (pp. 371-377).

[11] Rudra, K., Sharma, A., Bali, K., Choudhury, M., & Ganguly, N. (2019). Identifying and analyzing different aspects of English-Hindi code-switching in Twitter. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3), 1-28.

[12] Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020, December). Bi-LSTM and ensemble based bilingual sentiment analysis for a code-mixed Hindi-English social media text. In 2020 IEEE 17th India council international conference (INDICON) (pp. 1-6). IEEE.

[13] Anusha, M. D., & Shashirekha, H. L. (2021). BiLSTM-Sentiments Analysis in Code Mixed Dravidian Languages. In FIRE (Working Notes) (pp. 996-1004).

[14] Yadav, S., & Chakraborty, T. (2021, May). Zera-shot sentiment analysis for code-mixed data. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 18, pp. 15941-15942).

[15] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

[16] Lee, G. H., & Chen, Y. N. (2017). Muse: Modularizing unsupervised sense embeddings. arXiv preprint arXiv:1704.04601.

[17] Kusampudi, S. S. V., Chaluvadi, A., & Mamidi, R. (2021, September). Corpus creation and language identification in low-resource code-mixed Telugu-English text. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 744-752).

[18] Tang, T., Tang, X., & Yuan, T. (2020). Fine-tuning BERT for multi-label sentiment analysis in XLM-RoBERTa. International Journal of Advances in Soft Computing & Its Applications, 13(3)..

[19] Kannan, R. R., Rajalakshmi, R., & Kumar, L. (2021). IndicBERT Based Approach for Sentiment Analysis on Code-Mixed Tamil Tweets. In FIRE (Working Notes) (pp. 729-736).

[20] Sabri, N., Edalat, A., & Bahrak, B. (2021, March). Sentiment analysis of persian-english code-mixed texts. In 2021 26th International Computer Conference, Computer Society of Iran (CSICC) (pp. 1-4). IEEE.

[21] Wiciaputra, Y. K., Young, J. C., & Rusli, A. (2021). Bilingual Text Classification in English and Indonesian via Transfer Learning using XLM-RoBERTa. International Journal of Advances in Soft Computing & Its Applications, 13(3).

[22] Shanmugavadivel, K., Sathishkumar, V. E., Raja, S., Lingaiah, T. B., Neelakandan, S., & Subramanian, M. (2022). Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports, 12(1), 21557.

[23] Patil, A., Patwardhan, V., Phaltankar, A., Takawane, G., Joshi, R., 2023. Comparative study of pre-trained bert models for code-mixed hindienglish data, in: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), IEEE. pp. 1–7.

[24] Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., & Takalikar, M. S. (2023). Leveraging Language Identification to

Enhance Code-Mixed Text Classification. arXiv preprint arXiv:2306.04964.

[25] Patwardhan, V., Takawane, G., Kelkar, N., Gaikwad, O., Saraf, R., & Sonawane, S. (2023, March). Analysing The Sentiments Of Marathi-English Code-Mixed Social Media Data Using Machine Learning Techniques. In 2023 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-5). IEEE.

[26] Kumaresan, C., & Thangaraju, P. (2023). Elsa: Ensemble learning based sentiment analysis for diversified text. Measurement: Sensors, 25, 100663.

[27] Rathnayake, H., Sumanapala, J., Rukshani, R., & Ranathunga, S. (2022). Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowledge and Information Systems, 64(7), 1937-1966.

[28] Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Abidin, Z. Z. (2019). Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach. International Journal of Advanced Computer Science and Applications, 10(1).

[29] Fuadvy, M. J., & Ibrahim, R. (2019, October). Multilingual sentiment analysis on social media disaster data. In 2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE) (Vol. 6, pp. 269-272). IEEE.

[30] Mahadzir, N. H., Razak, N. H. A., & Omar, M. F. M. (2020, December). A New Sentiment Analysis Model for Mixed Language using Contextual Lexicon. In 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE) (pp. 1-5). IEEE.

[31] Tho, C., Heryadi, Y., Lukas, L., & Wibowo, A. (2021, April). Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach. In Journal of Physics: Conference Series (Vol. 1869, No. 1, p. 012084). IOP Publishing.

[32] Ahmad, T. M., & Abdullah, N. A. S. (2021). A Case Study on Social Media Analytics for Malaysia Budget. International Journal of Advanced Computer Science and Applications, 12(10).