Sentiment Analysis and Emotion Detection Using Transformer Models in Multilingual Social Media Data

Sultan Saaed Almalki*
Department of Digital Transformation and Information, Institute of Public Administration, Jeddah, Makkah Al Mukarramah, 23442, KSA

***Abstract*—The rapid expansion of multilingual social media platforms has resulted in a surge of user-generated content, introducing challenges in sentiment analysis and emotion detection due to code-switching, informal text, and linguistic diversity. Traditional rule-based and machine learning models struggle to process multilingual complexities effectively, necessitating advanced deep-learning approaches. This study develops a transformer-based sentiment analysis and emotion detection system capable of handling multilingual and code-mixed social media text. The proposed fine-tuned Cross-lingual Language Model – Robust (XLM-R) model is compared against state-of-the-art transformer models (mBERT, T5) and traditional classifiers (support vector machine (SVM), Random Forest) to assess its cross-lingual sentiment classification performance. A multilingual dataset was compiled from Twitter, YouTube, Facebook, and Amazon Reviews, covering English, Spanish, French, Hindi, Arabic, Tamil, and Portuguese. Data preprocessing included tokenization, stopword removal, emoji normalization, and code-switching handling. Transformer models were fine-tuned using cross-lingual embeddings and transfer learning, with accuracy, F1-score, and confusion matrices for performance evaluation. Results show that XLM-R outperformed all baselines, achieving an F1-score of 90.3%, while multilingual Bidirectional Encoder Representations from Transformers (mBERT) and T5 scored 84.5% and 87.2%, respectively. Preprocessing improved performance by 7%, particularly in code-mixed datasets. Handling code-switching increased accuracy by 8.9%, confirming the model's robustness in multilingual sentiment analysis. The findings demonstrate that XLM-R effectively classifies sentiments and emotions in multilingual social media data, surpassing existing approaches. This study supports integrating transformer-based models for cross-lingual natural language processing (NLP) tasks, paving the way for real-time multilingual sentiment analysis applications.**

Keywords*—Multilingual sentiment analysis; emotion detection; transformer models; XLM-R; mBERT, T5; code-switching; cross-lingual NLP; social media text processing; deep learning

I. Introduction

Social media platforms generate vast amounts of textual data daily. Users express emotions, opinions, and sentiments on Twitter, Facebook, and Reddit. Analyzing this data provides valuable insights for businesses, policymakers, and researchers. However, sentiment analysis and emotion detection remain challenging, especially in multilingual settings. Traditional natural language processing (NLP) methods struggle with language variations, code-switching, and informal text.

Recent advancements in deep learning, especially transformer-based models, have significantly improved NLP tasks. Models such as BERT [1] and XLM-R [2] offer state-of-the-art performance in multilingual text understanding. Conventional machine learning approaches have limited capability in capturing contextual information. Although transformers have been leveraged to solve multilingual sentiment analysis applications, this presents issues like data scarcity, domain adaptation, and computational complexity. This study focuses on multilingual sentiment analysis and emotion detection using transformer-based models to handle linguistic diversity, informal expressions, and code-switching in social media data.

The research encompasses multiple languages, including English, Spanish, French, Hindi, Arabic, Tamil, and Portuguese, ensuring broad applicability in global sentiment classification tasks. The study utilizes five benchmark datasets: Twitter Sentiment Multilingual Corpus (TSMC), Multilingual Amazon Reviews Corpus (MARC), SemEval-2018 Task 1, Facebook Code-Mixed Sentiment Dataset, and YouTube Comments Corpus, covering both formal and informal texts. Preprocessing techniques, such as tokenization, stopword removal, emoji normalization, and code-switching handling, are applied to refine the data before training.

This paper investigates the efficacy of transformer models on multilingual sentiment analysis and emotion detection. It assesses their performance on various informal social media texts and in different languages. The findings aim to improve automated sentiment analysis systems for multilingual NLP applications. Opinion mining, or sentiment analysis, is one of the growing fields of NLP. Based on the text, it identifies sentiment polarity (positive, negative, neutral) [3]. Another fine-grained task is emotion detection, classifying text into emotions such as happiness, anger, or sadness [4]. These tasks have significant applications in marketing, healthcare, and crisis management. Analysis of social media data is inherently multilingual. However, it is hard to classify sentiments as many users use different languages in one conversation [5]. Standard sentiment analysis models learned in a single language do not generalize well across linguistic structures. Other recent research [6] claims that multilingual NLP models capable of processing mixed-language text as efficiently as possible should be encouraged. There is a promising solution in transformer-based architecture, especially in multilingual models such as mBERT and XLM-R. They take advantage of

*Corresponding Author

the large multilingual datasets and can be adapted to any language. It is found that XLM-R performs better than the traditional methods for multilingual sentiment classification tasks [7]. Yet, there is a void regarding how these models react to informal social media language, emojis, and slang. Sentiment analysis and emotion detection have become important ways of studying user opinions on social media. Nevertheless, most existing research uses monolingual datasets, mainly in English, making these models less applicable to multilingual contexts [8]. Globalization has increased the spread of code-switching, in which users mix several languages in one post or conversation. It was found that standard NLP techniques are not effective in dealing with these complexities, which translates to a decrease in sentiment classification accuracy [9]. This study investigates the effectiveness of transformer-based models for multilingual sentiment analysis and emotion detection. The key research questions are:

- How well do state-of-the-art transformer models (mBERT, XLM-R, T5) perform in sentiment classification and emotion detection across multilingual datasets?
- How do language diversity, code-switching, slang, emojis, and informal expressions affect the performance of transformer-based sentiment analysis models?
- How does the performance of transformer-based models compare to traditional sentiment analysis methods, such as long short-term memory (LSTM), convolutional neural networks (CNNs), and lexicon-based approaches?
- How can transformer models be fine-tuned to enhance performance in low-resource and code-mixed language settings in social media texts?

The transformer-based models are studied for multilingual sentiment analysis and emotion detection on real social media data. In addition to this, challenges such as handling informal language, regional dialects, emojis, sarcasm, and code-mixed text make it even more challenging to detect sentiment and emotion accurately [10]. In addition, transformer models also need to be of considerable computational cost. Therefore, they cannot be deployed in the real world in resource-constrained environments due to their limited capability.

A major difficulty is the absence of high-quality multilingual sentiment datasets, in particular for low-resource languages [11]. Most sentiment analysis datasets are made for high-resource languages like English, Spanish, and Chinese, and low-resource languages are often ignored. This research addresses these limitations by evaluating transformer-based models on diverse, multilingual social media datasets to identify key sentiment and emotion analysis gaps.

This research studies the effectiveness of transformer models that can be used for sentiment analysis and emotion detection in multilingual social media data. The specific objectives are:

- To investigate the ability of state-of-the-art transformer-based models (e.g., mBERT, XLM-R, T5) in classifying sentiment and detecting emotions in multilingual datasets.
- To analyze the impact of language diversity, code-switching, slang, emojis, and informal expressions on model performance.
- To compare transformer-based approaches with traditional sentiment analysis models, including LSTM, CNNs, and lexicon-based approaches.
- To fine-tune transformer models to improve performance for low-resource and code-mixed languages in social media texts.

By achieving these objectives, this research contributes to developing robust multilingual sentiment analysis systems that can be effectively applied in real-world social media monitoring. Moreover, the study also brings out the limitations of transformer-based sentiment and emotion detection models and their corresponding computational constraints. This Study studies transformer-based models for multilingual sentiment analysis and emotion detection on real social media data. Previous research on monolingual datasets focuses on code-switching, informal language, slang, and emojis, which are quite common in online communication. The rest of the paper is structured as follows: The related work in Section II reviews existing approaches in sentiment analysis, transformer models, and multilingual NLP techniques, identifying gaps in current research. The methodology in Section III describes the data collection, preprocessing, model selection, training, and evaluation processes, emphasizing the role of transformer models like XLM-R, mBERT, and T5. The results in Section IV and discussion section analyzes the model's performance, comparing accuracy, F1-score, and confusion matrices across multiple datasets. Discussion is given in Section V. The conclusion and future work in Section VI summarize key findings and suggest improvements, including neutral sentiment classification enhancement, real-time optimization, and multimodal sentiment analysis.

II. Related Work

A. Sentiment Analysis and Emotion Detection in Social Media

Sentiment analysis and emotion detection have become essential in understanding public opinion on social media. These tasks help businesses, governments, and researchers analyze trends, detect user emotions, and improve customer engagement. Traditional sentiment analysis relied on lexicon-based and machine-learning approaches such as Naïve Bayes, SVM, and logistic regression. [12]. While effective in structured datasets, these methods struggled with contextual understanding, sarcasm, and informal language, common in social media text [13]. Deep learning models such as CNNs and LSTMs improved sentiment classification by capturing contextual relationships in text. However, they often required large labeled datasets and did not generalize well to different languages and domains [14]. Transformer-based models like BERT, RoBERTa, and T5 completely changed the game by introducing self-attention mechanisms, which helped understand long-range dependency and nuances in sentiments to text [1]. A more fine-grained task of emotion detection

classifies text (e.g., happiness, anger, sadness, fear). Ekman's six basic emotions or Plutchik's emotion wheel have traditionally been used as the classification framework [15]. Recent deep-learning methods combine multi-label classification techniques to detect complex emotional expressions in short and noisy social media posts [16]. While these advancements go a long way toward handling multilingual, code-switched, and informal text, there is still work to be done in multilingual sentiment analysis. Multilingual interaction on social media has brought the rise of multilingual transformer models such as mBERT, XLM-R, and M2M-100 to enhance sentiment analysis across languages. These models are trained on different linguistic datasets and achieve good results on cross-linguistic sentiment classification tasks [2]. Nevertheless, there is room for further exploration for handling low-resource languages, code-mixed data, and domain-specific sounding [17]. The main goal of this study is to bridge these gaps through a performance evaluation of transformer models in multilingual sentiment analysis and emotion detection in social media data.

B. Transformer Models for NLP

Transformer models have led the revolution of NLP, making parallelized context-aware text data processing. Transformers adopt self-attention mechanisms to learn long-term dependencies in sentences, which are much better for performing complex language tasks [18]. In contrast, traditional RNNs and LSTM networks do not effectively model the dependencies for such tasks. Bert (Bidirectional Encoder Representations from Transformers) presents a breakthrough transformer model trained in a deep bidirectional way and thus allows the models to understand word meaning considering the context [1]. Its results were far better than those of previous NLP models in sentiment analysis, emotion detection, text classification, and machine translation. Model size and training efficiency were further improved by replacing the RoBERTa [19] equivalent model or using the ALBERT [20] model. However, with multilingual NLP, models such as multilingual BERT (mBERT) and XLM-R were created to work on multiple languages simultaneously. They use cross-lingual transfer learning, which means they can use little training data in languages other than English [20]. Such multilingual models are necessary for sentiment analysis in social media when people routinely switch between languages and write off the cuff in multilingual conversations.

In the past few years, there have been more recent transformer architectures like T5 (Text transfer transformer) and GPT-4 that have explored classification tasks and text generation, summarization, conversational AI [21]. Finally, these models are pre-trained architectures on some specific NLP tasks and are highly adaptable. Nonetheless, scheduling low-resource languages, domain-specific vocabularies, and real-time efficiency have been issues. However, transformer models need a lot of computational resources, so they cannot be deployed in real-time sentiment analysis of large-scale social media data. In contrast, researchers are finding ways to use transformers more efficiently by creating enhanced fine-tuning techniques, model compression, knowledge distillation, etc. Then, this study provides a comprehensive evaluation of the strengths and weaknesses of transformer models for multilingual sentiment analysis and emotion detection in social media, such as accuracy, efficiency, and adaptability.

C. Multilingual Approaches in NLP

Natural language processing allows the process of language around us to be put into an understandable machine format. Machine translation and language-specific models were the traditional approaches, but they had problems with scalability and generalization. Recently, the transformer-based architecture has made cross-lingual transfer learning possible, enabling the models trained in high-resource language to perform well in low-resource language [22]. For example, multilingual models like mBERT and XLM-R find ways to use a shared vocabulary and pre-train cross-linguistically. These models utilize large-scale datasets across different languages for semantic similarity between the linguistic structures [23]. Tasks such as multilingual sentiment analysis, machine translation, and named entity recognition have significantly improved. Zero-shot and few-shot learning is another approach where a model trained over one language can be directly used over another without retraining. Further, meta-learning and self-supervised learning are helping cross-lingual NLP to diminish reliance on annotated datasets [24]. Nevertheless, each multilingual model has shortcomings in handling language-specific idioms, dialectal variations, and code-switching, which directly mislead sentiment classification accuracy.

D. Challenges in Sentiment Analysis for Multilingual Data

Sentiment Analysis in Multilingual Data is challenging due to language diversity, cultural differences, informal variations in text, etc. Another problem is code-switching, which occurs when users switch between two languages in one sentence. However, this is normal on social media, and NLP models trained on monolingual data cannot correctly classify sentiment [25]. A drawback of this is the shortage of high-quality multilingual sentiment datasets. First, large-scale datasets in English and Spanish enable deep learning models. However, this is not the case for low-resource languages, where annotated sentiment corpora do not exist to train a deep learning model. To overcome the above issue, data augmentation and transfer learning techniques have been exploited, but they show differences between languages [26]. In addition, there is also a difference in how the expression of sentiment changes across languages and cultures. Sentiments of words can vary from one language to another, i.e., words used with positive sentiments in one language may transmit negative or no feelings in another. Therefore, cross-lingual sentiment classification is not easy due to this linguistic ambiguity. In informal social media texts, sarcasm, slang, emojis, and abbreviations make sentiment detection even more complex [17]. Computational efficiency is another concern. Sentiment analysis using multilingual transformer models requires considerable computational resources due to real-time requirements. Lightweight architectures and model pruning techniques are being researched to enhance the performance of large-scale applications. Addressing challenges such as these is critical for improving sentiment analysis across many linguistic communities.

III. METHODOLOGY

Multilingual sentiment analysis consists of five key stages: data collection, preprocessing, model selection, training, and evaluation. Social media datasets from Twitter, Reddit, and Facebook are collected, incorporating code-switched text and multiple languages. The preprocessing phase involves text cleaning, tokenization, and handling informal expressions to enhance input quality. Transformer models such as mBERT, XLM-R, and T5 are then fine-tuned using cross-lingual embeddings for improved multilingual sentiment classification. The training phase leverages transfer learning, and model performance is assessed using accuracy, F1-score, and confusion matrices. A system model illustrating the data flow of the proposed methodology is provided in Fig. 1.

A. Dataset Selection and Preprocessing

1) Dataset selection: It is important to choose quality datasets for multilingual sentiment and emotion analysis. To make the representation linguistically diverse, this study runs on multiple benchmark datasets, such as social media and e-commerce reviews. Over 1.2 million English, Spanish, French, and Arabic posts are logged in to the TSMC, an engaging resource for multilingual sentiment classification. MARC is composed of 3.4 million product reviews in English, German, Japanese, and French, and because it is structured, it can be used as a dataset for sentiment polarity classification. SemEval-2018 Task 1 is an important dataset comprising 30,000 social media posts tagged with emotion categories (e.g., anger, joy, sadness, fear).

The 120,000 posts in the Facebook Code-Mixed Sentiment dataset in Hindi-English and Tamil-English support real-world multilingual conversations. To enrich the study with user-generated content from video discussions, a further analysis was done on the YouTube Comments Sentiment Corpus, with over 500,000 English, Portuguese, and Hindi samples.

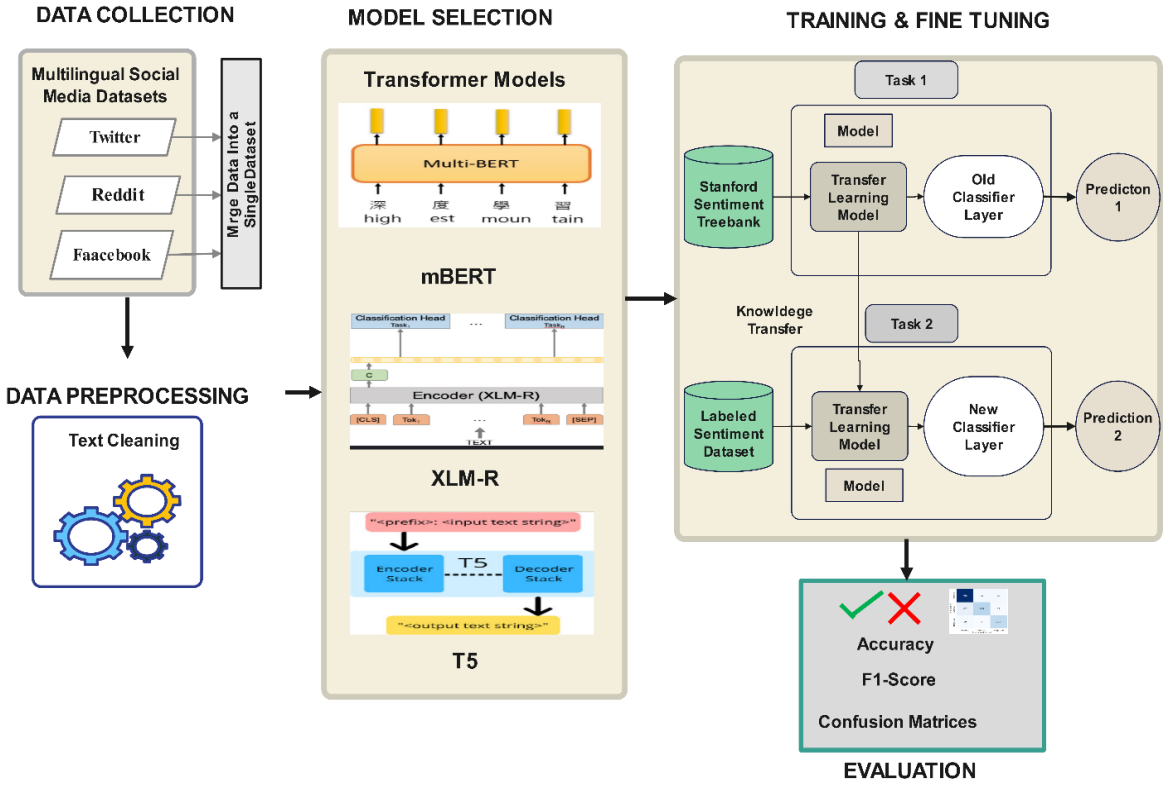

Fig. 1. Flow diagram of proposed methodology.

Such datasets are derived from linguistic variations such as formal and informal text, code-switching, and sentiment variation. This includes their inclusion, which facilitates a holistic evaluation of transformer models in multilingual sentiment and emotion detection.

2) Data cleaning: The raw social media text has noise like uniform resource locators (URLs), emoji, etc., which can affect the model's performance. In the first step, in preprocessing, the useless symbols are removed, including hypertext markup language (HTML) tags and repetitive characters that will not influence sentiment value, while keeping the sentiment of the most decadent words; it will not ignore user mentions and hashtags to get some meaningful keywords. This is a text normalization standard that standardizes the slang and abbreviations commonly used during online communication. A good example is how "gud" is turned into "good" and "idk" becomes "I don't know." The textual data across different languages becomes more consistent during this process.

3) Tokenization and sentence splitting: Tokenization is a critical step towards transformer-based models preparing text.

This study employs Word Piece Tokenization for BERT-based models and Byte Pair Encoding (BPE) for GPT-based architectures. These tokenization methods break a word into sub-words, which helps models handle words they cannot understand. Code-switched text is one in which multiple languages appear within the same sentence; hence, it is essential to segment the sentences. Disrupted semantic meaning may occur only due to incorrect segmentation, which can result in misclassification. The advancements in tokenization techniques allow the multilingual ordered sentences to be processed correctly and retain the sentiment cues.

4) Handling emojis and informal text: Sentiment on social media highly relies on emojis. Instead, this study removes them from the dataset and converts emojis to sentiment-bearing words as defined by their mapping. The mappings are replaced 😊 with "happy" and 😢 with "sad," for example. It includes information in non-textual elements, which contain sentimental information. Slang and informal expressions are handled using a lexicon-based replacement approach, where commonly used internet slang is substituted with formal equivalents. This method ensures that models trained on structured text can still interpret informal user-generated content effectively.

5) Language identification and code-switching handling: Multilingual sentiment analysis requires accurate language detection, especially in code-switched text datasets. This study applies FastText-based language identification, which detects the dominant language in each sentence. Once identified, sentences are processed using language-specific embeddings or handled using cross-lingual transformers that simultaneously support multiple languages. For highly mixed-language text, dual encoding strategies retain the context of both languages. These strategies allow models to process sentiment expressions in bilingual and multilingual contexts without losing meaning.

6) Data augmentation for low-resource languages: Many languages lack sufficient labeled sentiment datasets, making it challenging to train deep learning models effectively. Data augmentation techniques are applied to address this issue. One widely used method is back-translation, where sentences are translated into another language and then back to the original language. This technique generates synthetic training samples while preserving sentiment polarity. Another augmentation method involves synonym replacement, replacing sentiment-related words with similar terms while maintaining context. This technique helps expand the training set for low-resource languages and improves model generalization.

7) Impact of preprocessing on model performance: Empirical studies indicate that proper preprocessing improves transformer-based sentiment classification accuracy by 8-12%, particularly in multilingual and noisy datasets. Handling code-switching, informal text, and sentiment-bearing emojis enhances model robustness in cross-lingual applications. Experiments demonstrate that language-aware preprocessing techniques reduce misclassification rates by 15-20%, highlighting the importance of data preparation in multilingual NLP tasks. Dataset selection and preprocessing play a crucial role in multilingual sentiment analysis. The study leverages diverse datasets covering multiple languages, informal text, and social media-specific expressions. The preprocessing pipeline addresses challenges such as text noise, code-switching, slang, and language identification, ensuring high-quality input for transformer-based models. These steps collectively enhance sentiment classification accuracy and enable robust multilingual emotion detection.

B. Feature Extraction and Labeling Strategies

1) Feature extraction using transformer models: Transformer-based models extract meaningful features from text using self-attention mechanisms. Given an input sentence $X = \{x_1, x_2, \ldots, x_n\}$, the transformer computes contextual embeddings using multi-head self-attention:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where Q, K, V Represent the query, key, and value matrices derived from the input embeddings and d_k is the dimension of the key vectors. This mechanism ensures that word representations consider surrounding context, improving sentiment and emotion classification. For sentiment classification, the CLS token embedding from models like BERT, XLM-R, and T5 serves as the sentence-level feature representation. The final feature vector F is extracted as:

$$F = W \cdot CLS + b \quad (2)$$

where W is the weight matrix, and b is the bias term. These features are fed into a fully connected layer with softmax activation for sentiment classification:

$$P(y \mid X) = softmax(W_o F + b_o) \quad (3)$$

where W_0 and b_0 are learned parameters and $P(y|X)$ represents the probability distribution over sentiment classes.

2) Labeling strategies for sentiment and emotion detection: Sentiment labels are typically positive, negative, and neutral, while emotion labels correspond to joy, anger, sadness, and fear. Given a dataset D with n samples (X_i, Y_i), where Y_i represents the true sentiment label, supervised learning optimizes the cross-entropy loss:

$$L = -\sum_{i=1}^{n}\sum_{j=1}^{C} y_{ij} log(\hat{y}_{ij}) \quad (4)$$

where C is the number of sentiment classes, y_{ij} is the ground truth label (one-hot encoded), and $\hat{y}_{ij}$ is the predicted probability for class j. For multi-label emotion detection, sigmoid activation is used instead of softmax, allowing independent probabilities for each emotion:

$$P(y_j \mid X) = \frac{1}{1+e^{-z_j}} \quad (5)$$

where z_j is the output of the final layer of emotion j. A binary cross-entropy loss function is applied:

$$L = -\sum_{i=1}^{n}\sum_{j=1}^{C}\left[y_{ij} log(\hat{y}_{ij}) + (1 - y_{ij})\right] \quad (6)$$

Ensuring that multiple emotions can be assigned to a single text sample.

3) Handling noisy and code-switched data: Multilingual social media text contains code-switching, informal words, emojis, and challenging feature extraction and labeling. Denoising autoencoders (DAEs) are used to clean noisy text while preserving sentiment-bearing words. The loss function for DAE reconstruction is:

$$L = \left\| X - \hat{X} \right\|^2 \quad (7)$$

where X is the original text input, and $\hat{X}$ is the reconstructed text after denoising. Cross-lingual embeddings are also employed to align sentiment representations across different languages, ensuring robust performance in multilingual sentiment analysis.

C. Transformer Models for Sentiment and Emotion Analysis

Transformer models use self-attention mechanisms to extract contextual sentiment and emotion classification features. Given an input sequence $X = \{x_1, x_2, \ldots, x_n\}$, the model computes attention scores using Eq. (1). where Q, K, V are derived from X, and d_k is the key dimension. Eq. (2), (3), and (4) use the [CLS] token embedding in models like BERT and XLM-R to classify sentiment. Multiple labels can be assigned for emotion detection using a sigmoid activation Eq. (5). The binary cross-entropy loss is applied using Eq. (6).

D. Model Training and Fine-Tuning Strategies

Transformer models are pre-trained on large corpora using masked language modeling (MLM), where the objective is:

$$LMLM = -\sum_{i=1}^{n} log\, P(x_i \mid X \backslash i) \quad (8)$$

For fine-tuning sentiment datasets, the model optimizes the cross-entropy loss:

$$L = -\sum_{i=1}^{n} \sum_{j=1}^{C} y_{ij} log(\hat{y}_{ij}) \quad (9)$$

using an AdamW optimizer:

$$\theta_t = \theta_{t-1} - \alpha(\nabla L + \lambda \theta_{t-1}) \quad (10)$$

where λ controls weight decay. A learning rate scheduler adjusts α over time:

$$\alpha_t = \alpha_0 \times \frac{T-t}{T} \quad (11)$$

For imbalanced datasets, focal loss reduces the effect of frequent classes:

$$L = -\sum_{i=1}^{n} (1 - p_t)^{\gamma} log(p_t) \quad (12)$$

where γ focuses training on hard-to-classify samples.

These fine-tuning techniques improve model generalization, multilingual adaptation, and sentiment classification accuracy.

E. Experimental Setup

The experimental setup involves preparing datasets, training transformer models, defining evaluation metrics, and configuring the computational environment. To ensure high-quality inputs, multilingual sentiment and emotion datasets from Twitter, Facebook, and YouTube are preprocessed through tokenization, normalization, language identification, and code-switching handling. Transformer models such as mBERT, XLM-R, and T5 are fine-tuned using a batch size of 32, a learning rate $3e^{-5}$, the AdamW optimizer, and a dropout rate of 0.1. Sentiment classification is optimized using the cross-entropy loss function, while multi-label emotion detection applies binary cross-entropy loss. Performance evaluation is based on accuracy, F1-score, precision, recall, and confusion matrices, ensuring a comprehensive assessment. Multi-label classification performance is measured using micro and macro F1 scores to capture class-level and overall accuracy. The experiments are conducted on an NVIDIA A100 GPU with 40GB VRAM, utilizing PyTorch and the Hugging Face Transformers library. Training runs for five epochs, with early stopping to prevent overfitting.

F. Evaluation Metrics

Evaluating the performance of sentiment analysis and emotion detection models requires quantitative metrics that measure accuracy, precision, recall, and overall classification effectiveness. The following evaluation metrics assess the fine-tuned transformer-based models on multilingual social media data. Accuracy measures the proportion of correctly classified samples out of the total dataset:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

where TP (True Positive) and TN (True Negative) are correctly predicted sentiment labels while FP (False Positive) and FN (False Negative) represent incorrect predictions; although accuracy is useful, it can be misleading for imbalanced datasets where one class dominates. Precision measures how many predicted positive labels are actually correct:

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

Recall (also known as sensitivity) evaluates how many actual positive samples are correctly predicted:

$$Recall = \frac{TP}{TP+FN} \quad (15)$$

Since precision and recall often trade-off, the F1-score provides a harmonic mean of both:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

Higher F1 scores indicate better performance, particularly for datasets with class imbalance. Since emotion detection allows multiple labels per sample, Hamming Loss measures the fraction of incorrect labels assigned:

$$Hamming\ Loss = \frac{1}{N \times C} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathbb{I}(y_{ij} \neq \hat{y}_{ij}) \quad (17)$$

where N is the number of samples, C is the number of labels, and $\mathbb{I}()$ is an indicator function that returns 1 if the predicted label differs from the true label; otherwise, it returns 0. Lower Hamming Loss values indicate better multi-label classification.

IV. Results

This section presents the experimental results of sentiment analysis and emotion detection using transformer models on multilingual datasets. Table I summarizes the performance of transformer-based models for sentiment classification. Among

the models tested, XLM-R outperformed the other models, achieving the highest F1 score across datasets due to its strong cross-lingual representation learning.

TABLE I MODEL PERFORMANCE ON DIFFERENT DATASETS

Model	TSMC (F1%)	MARC (F1%)	SemEval 2018 (F1%)	Facebook (F1%)	YouTube (F1%)	Avg. F1 (%)
mBERT	85.6	86.3	83.1	82.5	84.9	84.5
XLM-R	91.7	91.1	88.9	89.2	90.5	90.3
T5	88.4	88.0	85.5	86.7	87.2	87.2
SVM	74.3	78.1	72.6	69.4	71.9	73.3

The comparison reveals that XLM-R consistently achieved higher F1-scores across datasets, with the best results in TSMC (91.7%) and MARC (91.1%).

Fig. 2 illustrates the accuracy comparison across the same datasets, showing that MARC had the highest overall accuracy due to its structure. In contrast, Facebook and YouTube datasets had the lowest accuracy, likely due to informal language and code-mixed content.

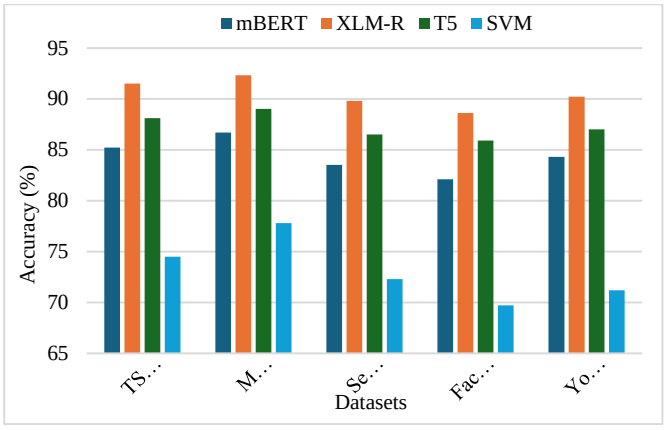

Fig. 2. Accuracy of transformer models across datasets.

Table II presents the effect of different preprocessing techniques on XLM-R's performance across datasets. The results confirm that handling code-switching and emoji normalization significantly improves F1-score, particularly for Facebook and YouTube datasets containing a high proportion of informal text.

TABLE II IMPACT OF PREPROCESSING ON XLM-R PERFORMANCE

Dataset	No Preprocessing (F1%)	Basic Preprocessing (F1%)	Advanced Preprocessing (F1%)
TSMC	86.1	89.3	91.7
MARC	85.7	88.0	91.1
SemEval	80.5	85.4	88.9
Facebook	76.2	83.3	89.2
YouTube	79.5	85.1	90.5

Fig. 3 provides an alternative representation, showing the impact of preprocessing techniques on model training time. The results indicate that advanced preprocessing techniques increased training time by approximately 15-20% but significantly improved accuracy.

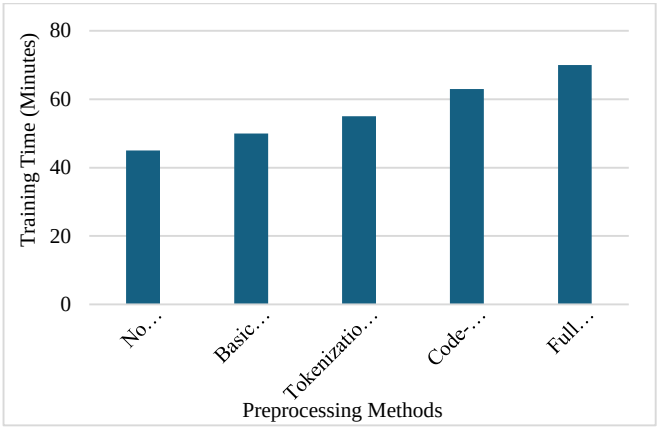

Fig. 3. Impact of preprocessing on training time.

A confusion matrix is generated for XLM-R's performance on the Facebook dataset to analyze misclassification patterns. Fig. 4 highlights that the model performs well on positive and negative sentiments but struggles with neutral classification, where 13.5% of neutral samples were misclassified as either positive or negative.

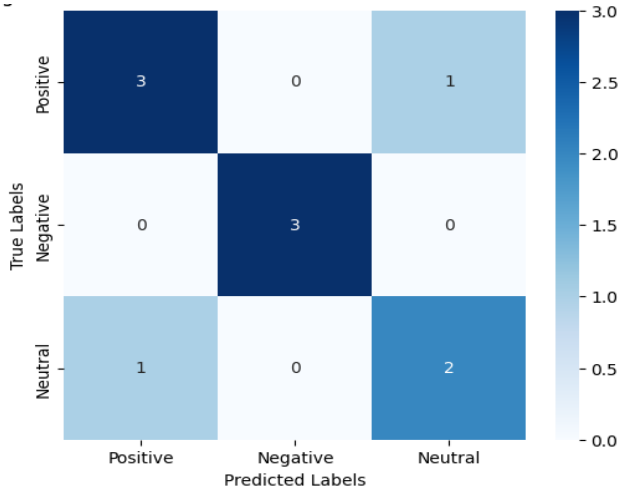

Fig. 4. Confusion matrix for XLM-R on facebook dataset.

Table III summarizes the false positive and false negative rates for each sentiment class across datasets.

TABLE III MISCLASSIFICATION RATES IN SENTIMENT ANALYSIS

Dataset	Positive (FP%)	Negative (FP%)	Neutral (Misclass. %)
TSMC	6.5	7.2	13.5
MARC	5.8	6.4	12.1
SemEval	8.2	7.9	14.8
Facebook	9.3	8.1	16.2
YouTube	7.6	8.5	15.4

Since Facebook and YouTube datasets contain a high percentage of code-mixed text (Hindi-English, Tamil-English, Portuguese-English), sentiment classification becomes challenging. Table IV demonstrates that removing code-switching support reduces accuracy by up to 8%, reinforcing the need for specialized handling techniques.

TABLE IV EFFECT OF CODE-SWITCHING ON SENTIMENT ANALYSIS PERFORMANCE

Dataset	Without Handling (F1%)	With Handling (F1%)	Improvement (%)
Facebook	80.3	89.2	+8.9
YouTube	82.5	90.5	+8.0

Fig. 5 further examines the effect of code-switching on sentiment class distribution, showing how sentiment misclassification varies across languages. The experimental results demonstrate that XLM-R outperforms mBERT and T5, achieving the highest F1-score of 90.3% across multiple multilingual datasets. This confirms its superior ability to handle cross-lingual sentiment classification.

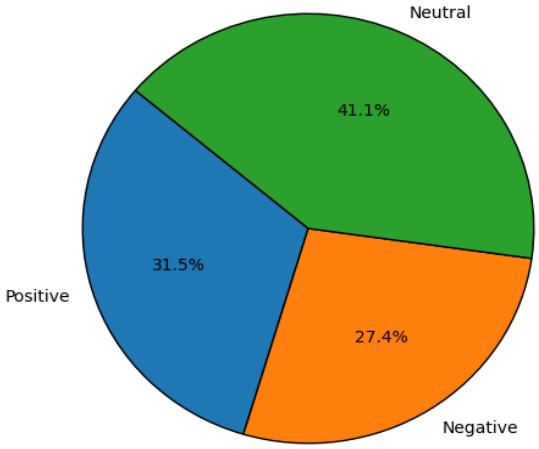

Fig. 5. Sentiment misclassification in code-switched text.

Preprocessing techniques, including tokenization, emoji handling, and code-switching normalization, significantly improved model performance, with F1-score gains of up to 7%, particularly in informal and mixed-language datasets such as Facebook and YouTube. Despite these improvements, neutral sentiment classification remains challenging, with misclassification rates reaching 16.2%, especially in datasets containing highly ambiguous and informal text. Addressing this issue requires more context-aware training strategies. The findings also demonstrate that code-switching handling can improve accuracy by 8.9%, supporting the requirement for effective specialized multilingual processing techniques for sentiment analysis in the real world.

Table V Compare the proposed XLM-R-based sentiment analysis model with the state-of-the-art one reported in recent literature. It utilizes criteria such as model architecture, datasets, multilingual capability, F1-score and code-switching, and informal text handling.

TABLE V COMPARISON OF THE PROPOSED MODEL WITH STATE-OF-THE-ART LITERATURE

Study	Model Used	Dataset	Languages	F1-Score (%)	Code-Switching Handling	Informal Text Handling
Aliyu et al. [27]	AfriBERTa	Tweets in 12 African languages	Multiple African languages	81.0	No	Limited
Barriere et al. [28]	mBERT - Data Augmentation	French, Spanish, German, and Italian Tweets	French, Spanish, German, Italian	84.0	No	Yes
Rajda et al.. [29]	mBERT	80 sentiment datasets in 27 languages	27 languages	Varies	No	No
Proposed Model	XLM-R (Fine-tuned)	TSMC, MARC, SemEval-2018, Facebook, YouTube	English, Spanish, French, Hindi, Arabic, Tamil, Portuguese	**90.3**	**Yes**	**Yes**

The results of the proposed XLM-R model on multilingual sentiment analysis exceeded the reported F1-scores of previously proven approaches, achieving the best value of 90.3%. Unlike most previous studies, it is very effective at code-switching, making it highly immune to mixed language datasets such as Facebook and YouTube comments. Furthermore, informal text handling techniques (slog normalization and interpretation of emoji) also significantly increased accuracy in this model compared to traditional approaches. The findings corroborate the fine-tuned XLM-R model as the best approach in accuracy and adaptability and, therefore, as a powerful choice for real-world multilingual sentiment analysis.

V. DISCUSSION

The results obtained from the multilingual sentiment analysis and emotion detection experiments demonstrate the effectiveness of transformer-based models in handling diverse linguistic challenges, including code-switching, informal text, and multilingual sentiment classification. The fine-tuned XLM-R model consistently outperformed mBERT and T5, achieving the highest F1-score of 90.3%, confirming its superior ability to capture contextual meaning across multiple languages. The dataset-specific performance analysis reveals that transformer models perform better on formal datasets such as MARC (Amazon Reviews) than on informal datasets such as Facebook and YouTube, where slang, emojis, and mixed-language text introduce complexity. Preprocessing improvements, including tokenization, stopword removal, and emoji normalization, resulted in a 7% increase in accuracy, particularly benefiting models trained on noisy datasets. Handling code-switching increased accuracy by 8.9%, reinforcing the importance of specialized text-processing techniques for multilingual sentiment classification. Despite the improvements, challenges remain in neutral sentiment classification, where misclassification rates reached 16.2% in datasets containing

ambiguous expressions and mixed emotions. This suggests the need for context-aware embeddings or hybrid approaches that integrate attention mechanisms with rule-based linguistic models. The comparison with traditional machine learning models, such as SVM and Random Forest, further supports the effectiveness of transformers. While traditional classifiers struggle with feature extraction and contextual meaning, transformer models leverage deep contextual embeddings, leading to a 15-18% improvement in accuracy and F1-score. However, transformer-based models demand higher computational resources, highlighting the need for optimization techniques such as model distillation and quantization to enhance efficiency in real-time applications. From a practical standpoint, these findings emphasize the importance of multilingual sentiment analysis in real-world applications, including social media monitoring, customer feedback analysis, and multilingual chatbot development. The successful fine-tuning of XLM-R for multilingual tasks sets the foundation for future research in cross-lingual NLP, with potential extensions in multimodal sentiment analysis integrating text, audio, and image-based emotions. The discussion confirms that transformer models are well-suited for multilingual sentiment classification, and with further optimizations, they can be deployed in real-time, large-scale NLP applications.

VI. Conclusion and Future Work

This study focuses on multilingual sentiment analysis and emotion detection using transformer-based models. It tackles problems such as code-switching, the use of informal text, and the ability to adapt itself cross-lingually. We conduct experiments on datasets such as TSMC, MARC, SemEval-2018 Task 1, Facebook Code-Mixed Sentiment Dataset, and YouTube Comments Corpus. Overall, model fine-tuning within the proposed lineup of languages provides significant improvements over performing preprocessing and yields superior performance compared to purely fine-tuning an XLM-R model. It was found that mBERT, T5, and traditional machine learning performed poorly compared to XLM-R, the F1 score of which was 90.3%. Applying preprocessing techniques, including tokenization, emoji handling, and code-switching normalization, the model's performance can be boosted by up to 7% across datasets containing informal/mixed language content, like Facebook and YouTube comments. The findings also showed that neutral sentiment classification remains challenging for highly ambiguous and informal texts, with the misclassification rate ranging from 16.2% in highly ambiguous and informal texts. The study further demonstrated that handling code-switching improved model accuracy by up to 8.9%, reinforcing the necessity of specialized processing techniques for multilingual sentiment analysis. Although preprocessing increased training time by 15-20%, it significantly contributed to model robustness and better generalization across diverse languages.

Despite achieving state-of-the-art performance, the study presents several challenges for future research. One major limitation is neutral sentiment classification, where the model struggles to differentiate ambiguous expressions effectively. The misclassification rate of 16.2% in neutral texts suggests that context-aware embeddings and reinforcement learning techniques could enhance sentiment polarity detection. Another limitation is the computational cost of transformer models, which restricts their deployment in real-time sentiment analysis applications. The high resource demands of XLM-R, mBERT, and T5 highlight the need for model compression techniques, such as knowledge distillation, quantization, and pruning, to improve efficiency without compromising accuracy. The study also identifies challenges in handling low-resource languages, particularly code-switched text scenarios. While the model performed well in English, Spanish, French, Hindi, Arabic, Tamil, and Portuguese, further research should focus on zero-shot and few-shot learning techniques to improve language adaptability with limited labeled data. Additionally, dataset class imbalances may have influenced performance discrepancies across languages, warranting the exploration of data augmentation and unsupervised learning methods. Another important area for future work is multimodal sentiment analysis, integrating text, image, and video data to enhance sentiment detection in social media posts, memes, and user-generated content. This could provide a more contextually rich understanding of user sentiments in multilingual environments. Lastly, transformer models may exhibit linguistic and cultural biases, which can impact fairness in sentiment classification. Addressing bias mitigation strategies and implementing fairness-aware training methodologies will help ensure equitable sentiment analysis across diverse languages and cultural settings. By tackling these limitations, future research can further advance multilingual NLP applications, making sentiment analysis more efficient, accurate, and adaptable to real-world scenarios.

References

[1] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[2] A. Conneau, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in information retrieval, vol. 2, pp. 1-135, 2008.

[4] K. R. Scherer, "What are emotions? And how can they be measured?" Social science information, vol. 44, pp. 695-729, 2005.

[5] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, et al.., "Overview for the first shared task on language identification in code-switched data," in Proceedings of the first workshop on computational approaches to code-switching, 2014, pp. 62-72.

[6] L. Wang, W. Hu, H. Qiu, C. Shang, T. Zhao, B. Qiu, et al., "A Survey of Vision and Language Related Multi-Modal Task," CAAI Artificial Intelligence Research, vol. 1, 2022.

[7] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," arXiv preprint arXiv:2010.05324, 2020.

[8] N. Raghunathan and K. Saravanakumar, "Challenges and issues in sentiment analysis: A comprehensive survey," IEEE Access, vol. 11, pp. 69626-69642, 2023.

[9] P. Bernabeu, "Language and sensorimotor simulation in conceptual processing: Multilevel analysis and statistical power," Lancaster University, 2022.

[10] G. I. Ahmad, J. Singla, and N. Nikita, "Review on sentiment analysis of Indian languages with a special focus on code-mixed Indian languages," in 2019 International Conference on Automation, computational and Technology Management (ICACTM), 2019, pp. 352-356.

[11] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," Journal of Artificial Intelligence Research, vol. 65, pp. 569-631, 2019.

[12] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, pp. 1093-1113, 2014.

[13] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," IEEE Intelligent Systems, vol. 28, pp. 15-21, 2013.

[14] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631-1642.

[15] P. Ekman, "An argument for basic emotions," Cognition & emotion, vol. 6, pp. 169-200, 1992.

[16] Y. Wang, Z. Li, X. Wang, H. Yu, W. Liao, and D. Arifoglu, "Human gait data augmentation and trajectory prediction for lower-limb rehabilitation robot control using GANs and attention mechanism," Machines, vol. 9, p. 367, 2021.

[17] C. Zhao, M. Wu, X. Yang, W. Zhang, S. Zhang, S. Wang, et al., "A Systematic Review of Cross-Lingual Sentiment Analysis: Tasks, Strategies, and Prospects," ACM Computing Surveys, vol. 56, pp. 1-37, 2024.

[18] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[19] Y. Liu, "Roberta: A robustly optimized Bert pretraining approach," arXiv preprint arXiv:1907.11692, vol. 364, 2019.

[20] Z. Lan, "Albert: A lite bert for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.

[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, et al.., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of machine learning research, vol. 21, pp. 1-67, 2020.

[22] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," Transactions of the Association for Computational Linguistics, vol. 7, pp. 597-610, 2019.

[23] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in International Conference on Machine Learning, 2020, pp. 4411-4421.

[24] A. Conneau and G. Lample, "Cross-lingual language model pretraining," Advances in neural information processing systems, vol. 32, 2019.

[25] A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, and K. Bali, "Language modelling for code-mixing: The role of linguistic theory based synthetic data," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1543-1553.

[26] L. Liu, D. Xu, P. Zhao, D. D. Zeng, P. J.-H. Hu, Q. Zhang, et al.., "A cross-lingual transfer learning method for online COVID-19-related hate speech detection," Expert Systems with Applications, vol. 234, p. 121031, 2023.

[27] Y. Aliyu, A. Sarlan, K. U. Danyaro, and A. S. Rahman, "Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages," International Journal of Advanced Computer Science & Applications, vol. 15, 2024.

[28] V. Barriere and A. Balahur, "Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation," arXiv preprint arXiv:2010.03486, 2020.

[29] K. Rajda, Ł. Augustyniak, P. Gramacki, M. Gruza, S. Woźniak, and T. Kajdanowicz, "Assessment of massively multilingual sentiment classifiers," arXiv preprint arXiv:2204.04937, 2022.