Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages

Yusuf Aliyu[1], Aliza Sarlan[2], Kamaluddeen Usman Danyaro[3], Abdulahi Sani B A Rahman[4]
Department of Computer and Information Science, Universiti Teknologi PETRONAS, Seri Iskandar 32610 Perak, Malaysia[1, 3, 4]
Center for Foundation Studies, Universiti Teknologi PETRONAS, Seri Iskandar, Perak, 32610, Malaysia[2]

***Abstract*—The analysis of sentiments expressed on social media platforms is a crucial tool for understanding user opinions and preferences. The large amount of the texts found on social media are mostly in different languages. However, the accuracy of sentiment analysis in these systems faces different challenges in multilingual low-resource settings. Recent advancements in deep learning transformer models have demonstrated superior performance compared to traditional machine learning techniques. The majority of preceding works are predominantly constructed on the foundation of monolingual languages. This study presents a comparative analysis that assesses the effectiveness of transformer models, for multilingual low-resource languages sentiment analysis. The study aims to improve the accuracy of the existing baseline performance in analyzing tweets written in 12 low-resource African languages. Four widely used start-of-the-art transformer models were employed. The experiment was carried out using standard datasets of tweets. The study showcases AfriBERTa as a robust performer, exhibiting superior sentiment analysis capabilities across diverse linguistic contexts. It outperformed the established benchmarks in both SemEval-2023 Task 12 and AfriSenti baseline. Our framework achieves remarkable results with an F1-score of 81% and an accuracy rate of 80.9%. This study provides validation of the framework's robustness in the domain of sentiment analysis across a low-resource linguistics context. our research not only contributes a comprehensive sentiment analysis framework for low-resource African languages but also charts a roadmap for future enhancements. Emphasize the ongoing pursuit of adaptability and robustness in sentiment analysis models for diverse linguistic landscapes.**

Keywords*—*Sentiment analysis; low-resource languages; multilingual, word-embedding, transformer

I. Introduction

In recent years, sentiment analysis has emerged as a pivotal research area in natural language processing [1]. It finds varied applications across various domains, including social media monitoring, customer feedback analysis, market research, and others [2] [3]. Sentiment analysis has been commonly classified into three levels based on various studies [4] [5] [3]. Thus, document-level sentiment classification focuses on discerning the overall sentiment expressed by an author in an opinionated text [6]. Sentence-level analysis, concentrating on individual sentences or arguments within a text. This is particularly valuable in subjectivity classification which assesses whether a sentence conveys an opinion [6] and lastly, the aspect level, is a more complex examination that aims to identify sentiments related to specific aspects of entities [7]. However, given the dynamic nature of social media. Sentence-level sentiment analysis finds particular significance in these platforms. It offers insights into user opinions and emotions on diverse topics. Moreover, X app formerly known as Twitter, has been a dynamic social media platform. It is a platform where users share a wide array of information in real-time [8]. It is considered one of the most vital sources for opinion mining and sentiment analysis [9]. Additionally, the X app contains a widespread multilingual text as individuals express their opinions in tweets spanning different languages, covering a variety of topics [10], [11].

The diversity in languages and cultures among users in the X app is evident in the multilingual nature of tweet texts [12]. This diversity opens up opportunities for businesses, governments, institutions, and other entities to find inferences about their entity for proper decision-making. Consequently, depending solely on sentiment analysis conducted in the English language carries a significant risk of missing crucial insights within written texts [13]. However, the increasing importance of multilingual sentiment analysis goes beyond high language limitations and becomes particularly relevant in low-resource languages.

The concept of low-resource languages encompasses various interpretations, including languages with limited research, facing resource scarcity, lacking computational support, being less commonly taught, or exhibiting low linguistic density [14]. In recent years, there has been a notable increase in the utilization of these languages on social media platforms, with a considerable proportion of users choosing to engage in communication through them [15]. Many languages in Africa and Asia fall into the low-resource category. They remain relatively unexplored in the realm of Natural Language Processing (NLP) research [16]. The underrepresentation of these languages in global economic, social, and political domains can potentially hinder economic and social progress. Nevertheless, advancing technology for low-resource languages can contribute to the increased participation of these language-speaking communities in the digital sphere [17]. Therefore, the abundance of multilingual content online underscores the necessity for sentiment analysis across various low-resource languages and cultures. Similarly, English remains the most extensively studied language [17], [18], [19], [20], [21]. Despite the linguistic diversity present in low-resource languages. The African languages with a substantial population received limited attention. Particularly in the context of sentiment analysis [22]. Low-resource languages pose unique challenges, including scarcity of labelled datasets [23], linguistic diversity, and limited computational resources.

Addressing sentiment analysis in such linguistic contexts necessitates a nuanced exploration of the capabilities and adaptability of transformer models.

The advent of transformer models, exemplified by architectures like BERT [24] (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), has considerably pushed the boundaries of sentiment analysis for diverse languages.

The study aims to enhance the accuracy performance of the current based-line performance on 12 African low-resource language tweets through rigorous hyper-parameter tuning and transformer comparison.

The paper is organized as follows: Section II gives a summary of related studies. Section III explores the proposed methodology for multilingual sentiment analysis within the framework. Section IV details the experiment and hyperparameter tuning. Section V showcases results and facilitates a discussion of the findings. Lastly, Section VI provides conclusions and outlines future directions for work.

II. Related Work

Sentiment analysis classification has different techniques, which are classified into three categories or classes lexicon-based, machine learning-based, and hybrid-based [9], [25], [26], [27], [28]. The machine learning techniques leverage well-known ML algorithms or models to address sentiment analysis. The models treat it as a typical text classification problem that incorporates syntactic and/or linguistic attributes [20]. Additionally, deep learning is a subset of the machine learning techniques [11], [29], based on the artificial neural network [7], [30]. The neural networks offer the most effective solutions for numerous challenges in image and speech recognition. As well as in the domain of natural language processing [31]. In recent years, deep learning models have shown remarkable performance in the field of sentiment analysis [28]. Especially transformer models. They are multilingual pre-trained based on transfer learning approaches that are trained on large various language text data [32], [33]. Moreover, the success of NLP has been primarily ascribed to transformers' capacity for learning broad language representations from enormous volumes of text input. Additionally, they have advanced to become the leading models in understanding and generating language [34] and It apply that knowledge learned to related tasks, producing astounding results [35]. However, prior studies utilized one or more of the above techniques for sentiment classification in one or more languages as:

The study of [13] performed sentiment analysis on English and Hausa tweets using an improved feature acquisition approach. They employed SVM, NB, and Maximum Entropy (MaxEnt) for classification. The experiment indicates that the classification models, when utilizing the aspect set, achieve a modest accuracy of 56% with the SVM classifier. In the case of the pure Hausa dataset SVM yields the best result. Similarly, [36] observed a scarcity of publicly accessible sentiment lexicons for low-resource languages, particularly notable in the Igbo language. In response, they advocated for the development of a comprehensive sentiment lexicon tailored for Igbo, designed to serve as a foundational resource for sentiment analysis in this linguistically underrepresented context. Another work by [37], performs feature learning and categorization using convolutional neural networks (CNN). They demonstrated the model's linguistic independence using the languages of English, French, and Greek with significant accuracy; however, it is highly domain-dependent as with only consists of a restaurant data set. So also, the authors [38], conduct a study that compares monolingual and cross-lingual sentiment analysis approaches utilizing the Hausa-English dataset. They translate most of the data using Google Translate before applying machine models. A similar approach was utilized by [39], the authors investigated the impact of translating from a language with ample resources to one with limited resources on emotion classification. They aim to address the observed gap leading to polarity changes and increase sentiment performance accuracy. Utilizing Google machine translation, they translate an English movie reviews dataset from IMDB into Urdu, Hindi, and German. Deep learning models with randomized parameters are employed and systematically adjusted to optimize results. The experimentation reveals varying accuracies, with English achieving the highest (88.37% via DNN), followed by Hindi (85.99% via Bi-LSTM), and Urdu (80.78%). Another work by [40] proposed a model that is independent of language for multi-class sentiment analysis, employing a straightforward neural network architecture. This is done on GenEval, Deutsche Bahn, and Arabic data sets with topic modelling. Their result shows that the deep neural network model can outperform traditional ML models when evaluated. In a different study presented by [41], find out that, there is a critique of the prevailing practice of constructing language models, particularly in low-resource scenarios. The inquiry revolves around the feasibility of translating data into English as an alternative, facilitating the use of pre-trained, comprehensive English language models. The researchers employ a contemporary machine translation approach to translate data into English that could potentially offer a solution to the challenges in multilingual sentiment analysis. Furthermore, their empirical experiment provides evidence that utilizing current baseline models on a huge scale does not result in a performance decrease, supporting the viability of this translation approach.

The work of [22] tries to bridge a gap of scarcity of data in low-resource language. They generated the inaugural extensive human-annotated Twitter sentiment dataset for four commonly spoken Nigerian languages, namely Hausa, Igbo, Pidgin, and Yoruba. In another research by [42], proposed a novel approach named AgglutiFiT, presenting an efficient fine-tuning strategy for pre-trained language models tailored for sentiment analysis and text classification. The fine-tuning process involves utilizing a low-noise dataset created through morphological analysis and stem extraction. The authors contend that this method excels in selecting pertinent semantic and syntactic information for low-resource languages like Kazakh, Kyrgyz, and Uyghur. Notably, the sentiment analysis task in Kazakh yielded an accuracy of approximately 92.87%. However, in research conducted by [43], analysed sentiment in code-mixed texts from users. They underscore the constrained predictability of conventional machine learning models

compared to deep learning counterparts using LSTM, CNN, and BiLSTM. Utilizing code-mixed data from Hindi-English and Bengali English, the experiment indicated that attention-based models outperformed traditional methods in accuracy by a margin of 20–60%. Additionally, when compared with monolingual English data, the accuracy reached 72.6% on the English monolingual dataset. In the study conducted by [44], the focus is on sentiment analysis of code-mixed Malaysian COVID-19-related news disseminated on Twitter. The researchers compile a multilingual Twitter dataset for COVID-19 encompassing tweets in Malay, English, and Chinese. Employing Byte-Pair Encoding (BPE) as a data compression technique, they apply two deep learning approaches: CNN and mBERT models. Results show that the BPE-M-BERT model exhibits a marginal performance advantage over the CNN model, emphasizing the advantageous adaptability of the pre-trained M-BERT network for a multilingual dataset. The researcher in study [45] investigates transformer fine-tuning methods for Hausa sentiment classification. Three pre-trained transformer multilingual language models, namely Roberta, XLM-R, and mBERT, are employed. The outcomes indicate that the mBERT-base-cased model achieves the best accuracy and F1-score, both reaching 0.73%.

The collective evidence from the diverse studies presented underscores the crucial importance of fine-tuning sentiment analysis models across multiple languages. Each study addresses specific linguistic nuances, cultural contexts, use of traditional ML models, and domain-specific challenges, emphasizing the need for a comprehensive approach to language diversity in sentiment analysis. From the creation of sentiment lexicons for underrepresented languages like Igbo, Hausa, Pidgin, and Yoruba to the exploration of code-switched data and the utilization of models such as XLM-R and mBERT across languages like Kazakh, Kyrgyz, Uyghur, Malay, Indian, and Chinese, these works collectively advocate for a nuanced understanding of sentiment in multilingual settings. Fine-tuning models across various languages allows for a more inclusive and accurate representation of sentiment expressions, reflecting the complex inherent in diverse linguistic structures and cultural contexts.

To address these limitations, the proposed framework emphasizes efficient parameter tuning to enhance model accuracy and performance. By optimizing model parameters, such as learning rates, batch sizes, and regularization techniques, the proposed framework aims to mitigate the shortcomings encountered in previous studies. Efficient parameter tuning enables the models to better capture the nuances of different languages and domains, thereby improving their overall effectiveness in sentiment analysis tasks. Table I present the summary of the related literature.

TABLE I. SUMMARY OF THE RELATED WORK

Study	Models/method	Language	limitation
[13]	SVM, MaxEn and NB	Hausa and English	Uses only traditional ML and the accuracy is below the benchmarking.
[36]	Manual data anotation	Igbo	Limited to Copus creation
[37]	CNN	English, French, and Greek	It highly domain dependent as with only consist of restaurant data set
[38]	NB, LSTM,BiLSTM, BERT, and Roberta.	Hausa	Depend on Google translator and domain specific
[39]	DNN, LSTM, Bi-LSTM and Conv1D	Urdu, Hindi, and German	Domain specific data, the model cannot be generalize to the domain.
[40]	ANN	Deutsche Bahn, and Arabic	The work is domain specific which rally on product data set
[22]	mBERT, Roberta,remBerta, XLR-R, and AfriBerta	Hausa, Igob, Yaruba, and Pidgen	Corpus creation for research progress and based-line evaluation. Need accuracy improvement
[41]	BERT, XLM-R	Scandinavian, Swedish, Norwegian Danish, Finnish	Depend on Machine translation which may not always be accurate.
[42]	XLM-R	Kazakh, Kyrgyz, and Uyghur	Different models may be explore for better performance.
[43]	BiLSTM, CNN, BERT	Hindi and Bengali	Different transformer models may be explore for evaluation.
[44]	CNN and mBERT	Malay, Indian, Chinese	Different transformer models may be explore for evaluation.
[45]	mBERT, Roberta and XLM-r	Hausa-English	Fine tuning specific to Hausa. It may not be generalize to other low resource language

III. METHODOLOGY

In this section, we provide an overview of the methodology framework for conducting sentiment analysis on tweets in low-resource multilingual settings. The pursuit of a better sentiment analysis framework tailored for multiple African languages tweets. This work adopts a distinctive approach that merges the strengths of transformer models through hyperparameter tuning. Specifically, use mBERT, Roberta, XLM-R, and AfriBERT from transformers. These models leverage the abundant semantic information present in tweets, offering a fundamental comprehension of word relationships through pre-trained embeddings. This approach facilitates the fine-tuning of models, thereby improving their adaptability to the distinctive linguistic nuances across various languages.

The core of our methodology involves the tweet datasets undergoing tokenization, feature detection, feature with model presentation, fine-tuning, training, and validating. To anticipate significant differences between these models, we conduct a rigorous statistical experiment. The decision-making process leads to the evaluation of model outputs using carefully chosen metrics. Fig. 1 serves as a visual guide, illustrating the proposed framework.

Through this framework, the objective is to contribute to a nuanced comprehension of sentiment in low-resource

languages and improve sentiment analysis accuracy. The development process of the sentiment classification models involves several distinct steps, commencing with the data description.

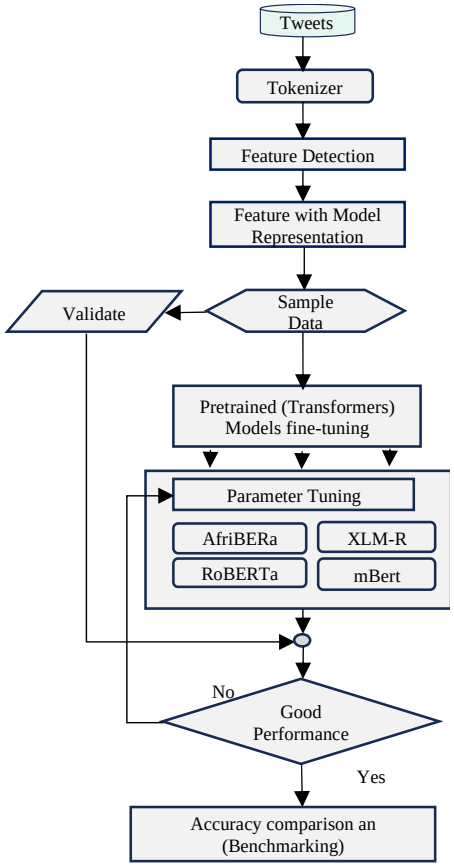


Fig. 1. Methodology framework workflow.

A. Multilingual Dataset

This stud utilized the Tweet data obtained from [46]. The data collection encompasses 12 distinct African languages, each characterized by unique linguistic features, writing systems, and language families, as outlined in Table II with their respective class label distribution. Spanning from Algerian Arabic, Moroccan Arabic/Darija, Hausa, Yoruba, Igbo, Nigerian Pidgin, Amharic, Swahili, Kinyarwanda, Twi, Mozambican Portuguese, and Xitsonga. The language and the class label distribution of the tweets are illustrated in Table II.

TABLE II. LANGUAGES AND THEIR RESPECTIVE CLASS DISTRIBUTION OF THE DATASETS

Language	Class label (distribution)		
	Positive	Neutral	Negative
Hausa	4687	4912	4573
Amharic	1332	3104	1548
Algerian Arabic	417	342	892
Darija Moroccan Arabic	1758	2161	1664
Swahili	1072	191	547
Yoruba	3542	3108	1872
Igbo	3084	4508	2600
Nigerian Pidgin	1808	72	3241
Xitsonga	384	136	284
Kinyarwanda	899	1257	1146
Twi	1644	522	1315
Mozambican Portuguese	681	1600	782

The datasets are provided as open-source resources explicitly crafted for research purposes, and they come pre-labelled. However, the training data comprises 63,685 instances of tweets, with 20,783 instances categorized as Positive, 20,108 as Negative, and 22,794 as Neutral. This distribution illustrates a fairly even representation of the three sentiment categories, reducing the potential for biased predictions towards any particular label. Fig. 2 provides a visual representation of the dataset distribution for additional reference. These datasets are subsequently employed to train feature detection and representation.

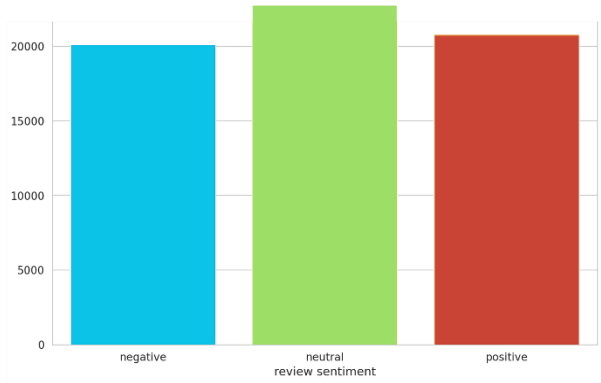


Fig. 2. Tweet class distribution.

B. Feature Detection

During the feature detection and extraction stage, textual data transforms into numerical form through tokenization. The Word Piece tokenizer is utilized in this procedure, fragmenting unfamiliar vocabulary words into sub-words, and consequently minimizing the occurrence of out-of-vocabulary words. This is accomplished through a greedy algorithm that prioritizes the longest possible match, thereby improving text-processing capabilities for transformers. Considering a tweet composed of N words, denoted as $T = \{w_1, w_2, w_3 \dots w_n\}$, each word w_i is processed and transformed into a numerical vector represented as e_i. Concurrently, the designated class y is converted into its vector, denoted as v_a.

$$e_i = E(w_i) \quad (1)$$

$$v_a = E(y) \quad (2)$$

The initialization of context embedding vectors, e_i, and class vectors, v_a, is described by Eq. (1) and Eq. (2). Eq. (1) denotes the transformation of each word, w_i, into a numerical vector, e_i, while Eq. (2) represents the conversion of class y into its vector v_a. In these equations, $E \in \mathbb{R}^{v \times d}$ denotes the embeddings, where d represents the dimension of the word embedding vectors. The vocabulary size is denoted as v, and N

represents the number of words in the tweet. Subsequently, the input vectors are passed into the transformers for fine-tuning.

C. Fine-turning

The fine-tuning process begins with the mBERT model transformer for sentiment classification, incorporating its pre-trained weights. Tokenized input sequences are enriched with a special classification token ([CLS]) at the beginning and a separation token ([SEP]) at the end. Token embeddings for each sub-word are generated using the embedding matrix and are combined with segment embeddings to differentiate between tokens from the first and second sentences. Position embeddings are employed to denote the position of each token in the input sequence. The BERT input representation, consisting of token embeddings, segment embeddings, and position embeddings, is then inputted into a SoftMax layer for classification aggregation. Fig. 3 depicts the visual representation of transformer architecture for the classification task.

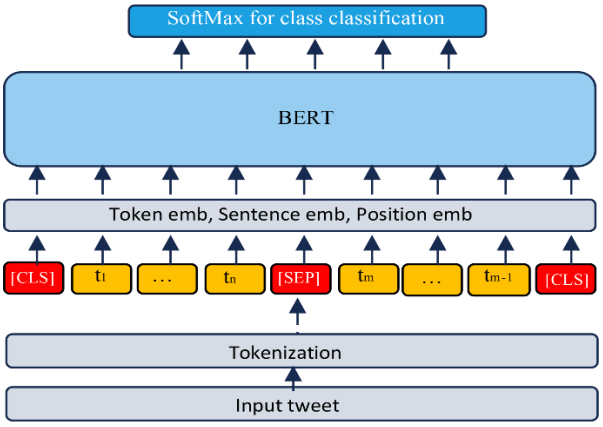


Fig. 3. Transformer model for classification.

Similarly, in the case of the Roberta model, we followed the same comparable fine-tuning approach, customizing the pre-trained weights specifically for sentiment analysis in low-resource contexts. The tokenization procedure, utilizing the Transformer tokenizer, and the creation of the input representation, which includes token embeddings, segment embeddings, and position embeddings, remained consistent with the models described earlier. In the end, a SoftMax layer was utilized for sentiment classification.

In addition to mBERT and Roberta, we included the XLM-R and afriBERTa models in the framework. These models adhered to a comparable methodology as the pre-trained mBERT models. The tokenized input sequences were converted into the input representation, which included token embeddings, segment embeddings, and position embeddings. Following this, the input representation underwent processing through a SoftMax layer for sentiment classification. The components of the model play a crucial role in the effective representation of language by transformers [24].

Furthermore, the models often require fixed-length input sequences for effective processing, a concept known as padding. The special token used for padding in transformers is generally the [PAD] token, which is inserted into the input sequence to fill the remaining empty spaces until the sequence reaches the desired fixed length. Additionally, any unknown token is represented as [UNK]. Tokens from the first sentence are assigned the label "0," while those from the second sentence are labelled "1". Position embeddings indicate the position of each token in the input sequence.

The selection of the aforementioned models was influenced by their proficiency in multilingualism, as discussed in the work by [47]. They are encompassed with some African languages. These chosen pre-trained language models were specifically designed to tackle this challenge and have demonstrated strong performance in multilingual settings, as highlighted by [48]. Given their robust nature, various hyperparameter combinations were employed to optimize their effectiveness in comprehending the complexities of different languages.

D. Hyperparameter Tuning

To determine optimal parameters, a series of experiments were conducted, employing different batch sizes, and assessing the performance of mBERT on a validation set. Table III provides a detailed presentation of results obtained from hyperparameter tuning. These values play a crucial role in the model optimization process. Each row in the table represents a unique combination of learning rates and batch sizes, depicting the mBERT model's performance in relation to the F1-score.

TABLE III. HYPERPARAMETER COMBINATION AND PERFORMANCE

Parameter						
	Learning rate	Batch size				Metric
sn		8	16	32	64	
1	5e-5	0.32	0.43	0.48	0.38	F1-score
2	3e-5	0.31	0.52	0.60	0.41	F1-score
3	2e-5	0.51	0.53	0.58	0.54	F1-score
5	1e-6	0.58	0.56	0.62	0.60	F1-score

The chosen hyperparameter values were systematically selected to assess their influence on the sentiment analysis tasks. Striving to pinpoint combinations that achieve an ideal equilibrium between model convergence and computational efficiency. The primary objective was to finds a configuration that strike a favorable equilibrium for the task. A detailed analysis of the table exposes discernible patterns. Particularly in terms of F1-score variations across different learning rates and batch sizes. The experiments were carefully carried out to measure the influence of hyperparameter selections on the overall performance of the model. Remarkably, it was observed that employing a learning rate of 1e-6 in conjunction with a batch size of 32 yielded the highest f1-score. This finding validated the model's ability to find a balance between efficient computation and robust model convergence. Subsequently, these values were adopted for the remaining experiments involving other models in the study. The moderate batch size has an advantage over the smaller batch which exhibits more variability in their performance as in the f1-score validation in Table III. On the other hand, a bigger batch size led to slower convergence of the training process which caused the model to overfit on the validation data. Therefore, utilizing moderate batch sizes offers several advantages, including

enhanced memory efficiency, accelerated convergence, and improved regularization effects by introducing sufficient noise to mitigate overfitting. Additionally, training with moderate batch sizes is observed to contribute to heightened stability in selecting an optimal learning rate as in Table III.

The selection of the maximum sequence length parameter is set at 150. As a result of an analysis of the tweet distribution within the datasets. This choice was made precisely to minimize the requirement for unnecessary padding during the training. This aims to capture the primary concepts conveyed in the tweets effectively. Fig. 4 visually depicts the maximum token length observed in tweets, illustrating the determined maximum sequence length for better context. The choice of a learning rate of 1e-6 was arrived at after a comprehensive examination of diverse parameter combinations throughout the training process, as outlined in Table III. This is a lower learning rate. A lower learning rate has demonstrated effectiveness across multiple Natural Language Processing. (NLP) tasks. Additionally, a systematic dropout tuning process was conducted, exploring a range of dropout rates from 0.5 to 0.3. This iterative experimentation aimed to pinpoint the dropout rate that strikes the best balance between mitigating overfitting and enhancing model generalization.

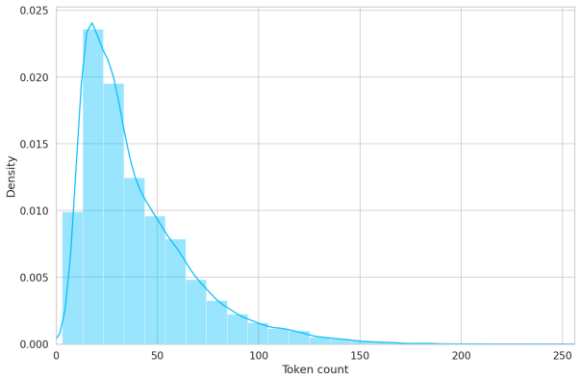


Fig. 4. Tweets token maximum sequence length.

Through this process, it was observed that a dropout rate of 0.3 consistently resulted in improved accuracy in this task compared to higher dropout rates. This careful tuning ensures that the chosen value aligns with best practices and is empirically grounded in its positive impact on model performance. For optimization, the widely employed Adam optimizer was utilized, known for its efficiency and effectiveness in training deep learning models. The same set of hyperparameter values was consistently applied across other models, as depicted in Table IV, outlining the chosen hyperparameters.

TABLE IV. HYPERPARAMETER USED

Sn	*Parameter*	*Value*
1	Max Sequence length	150
2	Batch size	32
3	Learning rate	1e-6
4	Optimizer	Adam
5	Dropout	0.3

IV. EXPERIMENT

This research explores the utilization of transformers' fine-tuning techniques for sentiment classification tasks in multilingual tweets. Specifically, it delves into the application of the Roberta, XLM-R, mBERT, and AfriBERT models in this context.

A. System Implementation

This research conducted experiments using the Python programming language, TensorFlow version 1.13.1, and the torch library version 2.0.1+cu118 for multilabel classification. The implementation took place on Google Collaboratory, utilizing a GPU hardware accelerator to enhance computational efficiency. Various tools, including NumPy, Pandas, sci-kit-learn, transformers, and seaborn libraries, among others, were employed to facilitate the analysis of tweets. The obtained results highlight the effectiveness of the implemented framework and emphasize the importance of adapting the code to specific requirements. The experiment notebook is accessible at: https://github.com/yusuf-003/Multilingual_experiment

B. Data Description

In this study, a dataset consisting of 63,685 tweets in 12 low-resource African languages was utilized. The dataset comprises three columns: id, tweet, and labels. The class labels are multiclass categorical data named as positive, negative, and neutral. To prepare the data for machine learning algorithms, the class labels were converted to numerical data using a dictionary-based class mapping function. This preprocessing step was conducted before feeding the data into the machine learning algorithm. Table V provides a sample of the dataset. The datasets are available at https://github.com/afrisenti-semeval/afrisent-semeval-2023/tree/main/SubtaskB

TABLE V. TWEETS SAMPLE

ID	*Tweet*	*Label*
mul_001	if i dey enter your eye or you like me and no fit talk am time dey go ohreporting live from paris	Positive
mul_002	@user @user Ndi igbo is na ara di na udi	Negative
mul_003	الا شوايةفوق من عندستة فكوم بـ يوكوم كاملين الـ فاخر هو بـ ذكـ يران	Negative
mul_004	SAMIA ATOA ANGALIZO KUIKABILI SARATANI Makamu wa Rais Samia Suluhu Hassan amesema tatizo la Saratani kwa Watanzania linaweza kupungua ama kuondokana nalo endapo kutakuwa na tabia ya.	Neutral
mul_005	Dùnùndún, ọ̀jọ̀jọ́, ____ Ẹ̀bà, ____, àmàlà #Ibeere #Yoruba	Neutral
mul_006	ኩላሊት፣ ጉበት፣ ልብ፣ ሳንባ እና ሌሎችም የሰውነት አካላት ከለጋሾች ሰውነት ከወጡ በኋላ ወደ ታካሚው ገላ እስከሚገቡ ድረስ ያለው ቆይታቸው ጥያቄን ይፈጥርብዎታል? በዚህ ፅሁፍ መልሱ…	Neutral
mul_007	@user Wai jihar shugaban kasa kenanfa ..ammafa dik'abundake faruwa yanaji .yakasa yinkumai...Allah ka yayemana musibannan🏘🏘	Negatvie
mul_008	في اصلاحات أقوى صاحب ملكنا أن: تعلم هل 'والـ شرق؟ الـ غرب بـ شهادة الـ عربـ ي الـ عالم	Positive

The dataset was randomly partitioned into three subsets for training purposes. These subsets were designated as train, validation, and test sets. 90% of the data was allocated for training, amounting to a finalized set of 57,316 tweets, while the validation set for 5% accounted for 3184. Lastly, 5% was separated for testing, resulting in 3185 tweets data. The importance of splitting the data into three parts is that it allows for a more robust evaluation of the model. The models were trained on the training data, and their performance are evaluated on the validation set. The hyperparameters of the model are tuned based on the performance of the validation set. Finally, the model was evaluated on the test set to estimate its performance on unseen data. Splitting the data into three parts helps to avoid overfitting and to have a better estimate of the model's generalization performance. The models are evaluated using the standard machine learning evaluation metric.

C. Evaluation Metric

Evaluation metrics are crucial instruments for assessing the performance of machine learning models. Through comparison of the predicted outcomes to actual results. Higher scores indicate superior predictive capabilities. It is a guiding parameter optimization for optimal performance during the tuning phase. Understanding the fundamental definition of metrics is crucial for alignment with system goals. Inaccurate evaluation using diverse metrics can lead to challenges in deploying systems on unobserved datasets. This may result in suboptimal predictions. The emphasis of this study is on Natural Language Processing metrics, which encompass precision, recall, F1-Score, and Accuracy. These metrics are derived from the confusion matrix, incorporating elements like true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) values.

1) Accuracy metric: To Accuracy is a metric that measures the frequency of correct sentiment ratings. The goal of accuracy testing is to showcase the efficiency of the suggested framework in predicting data. The accuracy formula is given as in Eq. (3):

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

2) Recall metric: Recall, essentially, assesses the outcome by determining how many of the genuinely relevant results are retrieved. The goal of recall testing is to appraise the proposed framework's ability to appropriately recall correctly classified data. The recall formula is given as in Eq. (4):

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3) Precision metric: Precision signifies the precision of the proposed framework in forecasting the accurate class for a particular tweet. This aspect underscores the model's capability to minimize false positives and ensure the correctness of predicted values corresponding to the designated class. Eq. (5) provides the formula:

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

4) F1-score metric: The f1-score represents the interplay between precision and recall, and it is inversely proportional to strike a balance between the two. Additionally, the f1 score serves as a harmonic mean [39]. It effectively mitigates the imbalance between precision and recall in the evaluation of a model's performance. The formula is given as in Eq. (6):

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

V. Result and Discussion

This section reveals the results of the experiment. It is rigorously analyzed, and thoughtfully presented. It involves a thorough investigation of the evaluation results, comparative metrics and conclusions drawn from the study. The comparison of performance evolution is grounded in results obtained from training and testing the unseen data. Initially, the training epoch is set at a maximum of 20 epochs. The experiment delves into various transformers, including mBERT, RoBERTa, XLM-R, and AfrBERTa.

A. Performance Evaluation

The effective evaluation of the proposed method in classifying tweets test data into sentiment categories: Positive, Negative, and Neutral. It encompassed several key models of varying multilingualism complexities. Table VI illustrates the results obtained from the experiment.

TABLE VI. Sentiment Analysis Model Performance Comparison and Evaluation Performance

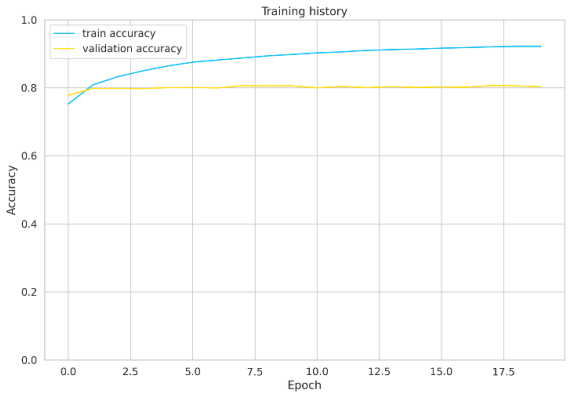
Model	Evaluation Metric			
	Accuracy	***Recall$_{ma}$***	***Precision$_{ma}$***	***F1-score***
mBERT	0.6229	0.62	0.63	0.62
RoBERTa	0.6556	0.75	0.75	**0.74**
XLM-R	0.6411	0.64	0.64	0.64
AfriBERTa	0.8088	0.81	0.81	**0.81***

The results of the experiment in Table VI offer a comprehensive insight into the performance of multilingual sentiment analysis models. The experiment was done across a diverse range of low-resource African languages. The evaluation metric. Thus accuracy, recall (macro-average), precision (macro-average), and F1-score offer a nuanced comprehension of the capabilities and constraints of each model.

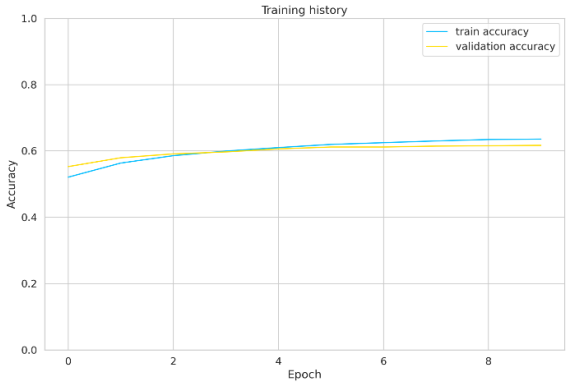
AfriBERTa stands out as the most effective model, achieving an impressive accuracy of 0.8088 and a balanced F1-score of 0.81. The model's specialization for African languages, coupled with fine-tuning on a diverse linguistic dataset, appears to be a key factor in its superior performance. This allows AfriBERTa to capture complex linguistic patterns and cultural context, essential for accurate sentiment analysis in this context. RoBERTa demonstrates commendable performance, securing the second-highest accuracy (0.6556) and a competitive F1-score of 0.74. Its general-purpose nature enables adaptability across different languages, outperforming mBERT and XLM-R. Although RoBERTa doesn't surpass AfriBERTa, recall and precision values of 0.75 indicate robustness, albeit slightly below the specialized model.

The decision to fine-tune models across multiple languages proves advantageous, particularly evident in the success of AfriBERTa. This approach enhances the model's robustness, allowing it to effectively handle the linguistic diversity present in African tweets. The contrast in performance metrics underlines the importance of considering regional linguistic variations when developing sentiment analysis models. AfriBERTa not only achieves high accuracy but also strikes a balance between precision and recall. A precision of 0.81 indicates that when AfriBERTa predicts a sentiment, it is highly likely to be correct. Simultaneously, a recall of 0.81 suggests the model captures a substantial portion of positive, negative, and neutral sentiments, crucial for an understanding of sentiment distribution.

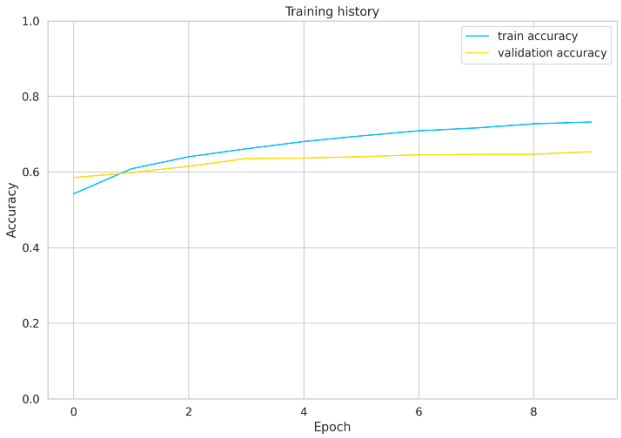
Fig. 5 illustrates the graphical representations of the model's train and validation accuracy curve of the experiment. However, the models' curves, indicate a steady improvement in accuracy as training progresses. This suggests that all the models effectively learned from the data. Moreover, an observation is the noticeable difference between the accuracy scores on the training and validation datasets, especially in the first to third training epoch in Fig. 5(a) and Fig. 5(c). This discrepancy implies the possibility of overfitting in the Afribarta and Robarta models, where the model may be overly tailored to the training data. This may be course due to variations in data distribution across languages. Although, the Afriberta model in Fig. 5(a), exhibits substantial overfitting despite its effectiveness in performance. While Fig. 5(b) mBart and Fig. 5(d) XLM-r control, this issue suggests that its pretraining on a diverse dataset might have played a crucial role in enhancing its ability to generalize effectively.



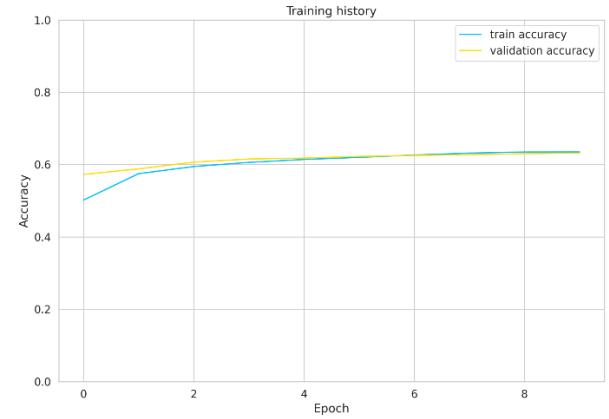
(a)



(b)



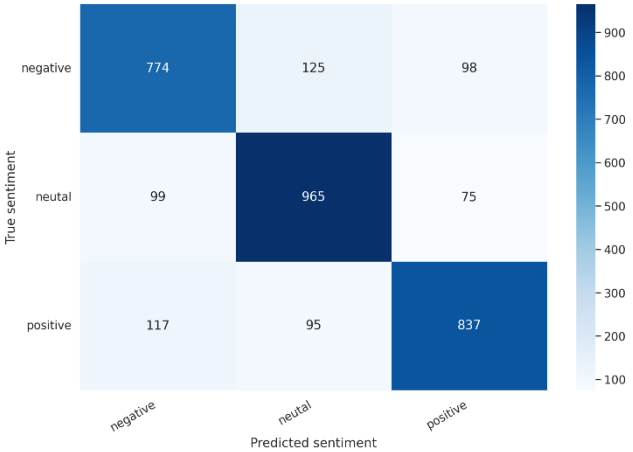
(c)



(d)

Fig. 5. (a) AfriBerta. (b) mBert. (c) Roberta. (d) XLM-r.

Similarly, Fig. 6 illustrates the confusion matrices, depicting correct and incorrect predictions by the models. Thus, a noticeable disparity is observed, with the Fig. 6(b) mBart, Fig. 6(c) RoBERTa, and Fig. 6(d) XLM-r exhibiting a higher rate of misclassifications compared to Fig. 6(a) Afriberta, which demonstrates a small number of incorrect predictions. Afriberta displays exceptional performance in accurately classifying each distinct category.



(a)

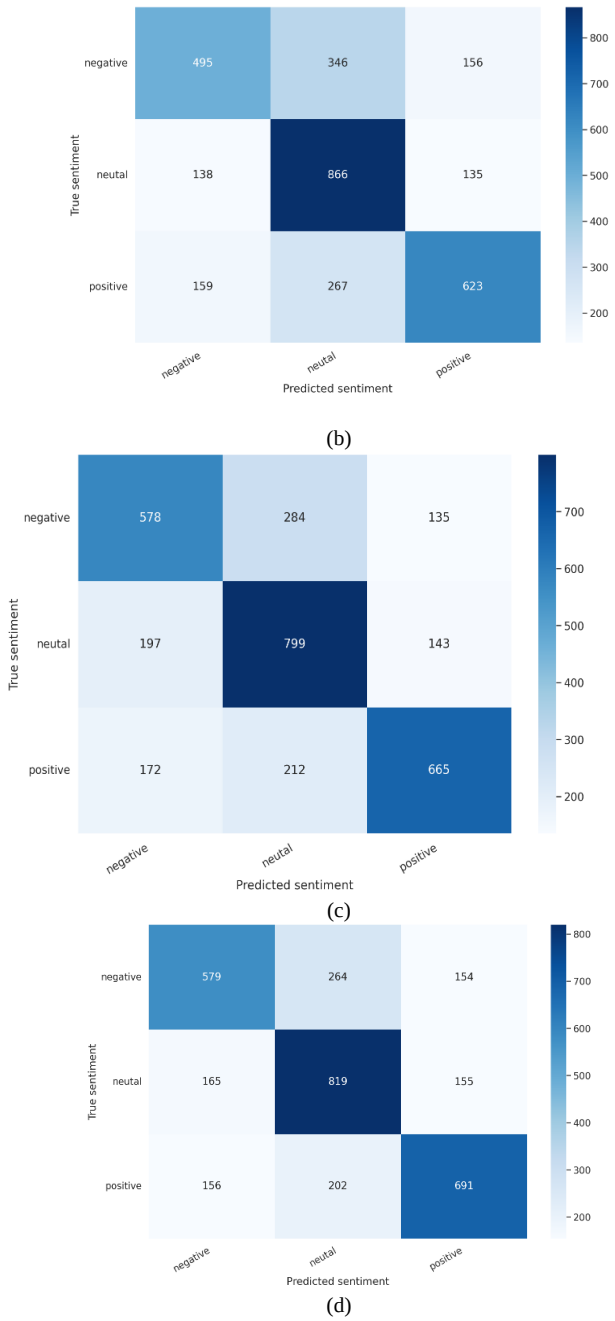


Fig. 6. (a) Afriberta. (b) mBert. (c) Robarta, (d) XLM-r.

B. Error analysis Evaluation

The training and validation metrics presented in Table VII provide a detailed view of the performance of each model during the training process.

TABLE VII. : SENTIMENT ANALYSIS MODEL PERFORMANCE COMPARISON AND EVALUATION PERFORMANCE

Model	Training metric			
	$Training_{loss}$	*Acc*	$Validation_{loss}$	*Acc*
mBERT	0.9037	0.6354	0.9189	0.6165
RoBERTa	0.8137	0.7322	0.8907	0.6536
XLM-R	0.9023	0.6349	0.9078	0.6319
AfriBERTa	0.6619	0.8874	0.7429	0.8059

Analyzing the values offers insights into potential sources of errors and the model's capacity to generalize to unfamiliar data. mBERT and XLM-r, relatively shows low training loss and exhibit a noticeable drop in accuracy during validation. The validation accuracy suggests that the model struggles to generalize to new data, possibly due to an excessive focus on training data specifics.

AfriBerta and RoBERTa demonstrate a more balanced performance with training and validation accuracy. The Roberta model exhibits a slightly higher training loss (0.8137), suggesting a degree of complexity in the learned representations. The marginal decrease in accuracy during validation might be attributed to the model's challenge in capturing subtle patterns in the validation set. The AfriBERTa exhibits impressive training accuracy (0.8874) and validation accuracy (0.8059), indicating robust learning and generalization capabilities. The comparatively low training loss (0.6619) further supports the model's capability to identify complex patterns within the data. AfriBERTa's performance suggests effective fine-tuning across multiple languages. It contributing to its superior performance in sentiment analysis as observed in the earlier sections. Fig. 7 illustrates the visual model train/accuracy and loss comparison.

C. State-of-the-art Benchmarking

To evaluate the efficiency of the proposed framework, a comprehensive evaluation was undertaken, where in the performance was compared against the best-reported results within the realm of sentiment analysis for low-resource languages. Notably, the framework demonstrates superior performance across multiple benchmarks, surpassing models that currently lead in the domain.

In the SemEval-2023 Task 12: Sentiment Analysis for African Language [46] a renowned competition in the field, the best-reported F1-score for multilingual task was 75.06%. Remarkably, the proposed framework surpasses this benchmark, achieving an impressive F1 score of 80.6%, showcasing its robustness in capturing sentiment in multilingual tweets.

Furthermore, the framework was compared against the AfriSenti benchmark [49]. A Twitter Sentiment Analysis benchmark for African Languages, where the reported best F1 score stands at 71.2%. In comparison, the framework exceeds this benchmark, highlighting its ability to handle the complexities of sentiment analysis in a multilingual context.

The model also outperforms NaijaSenti [22], a Nigerian Twitter Sentiment Corpus, including four languages in our training data. It achieved an average F1-score of 78.3%. The robustness of the framework is further emphasized by its ability to surpass these benchmarks, achieving an outstanding F1 score of 80.6%.

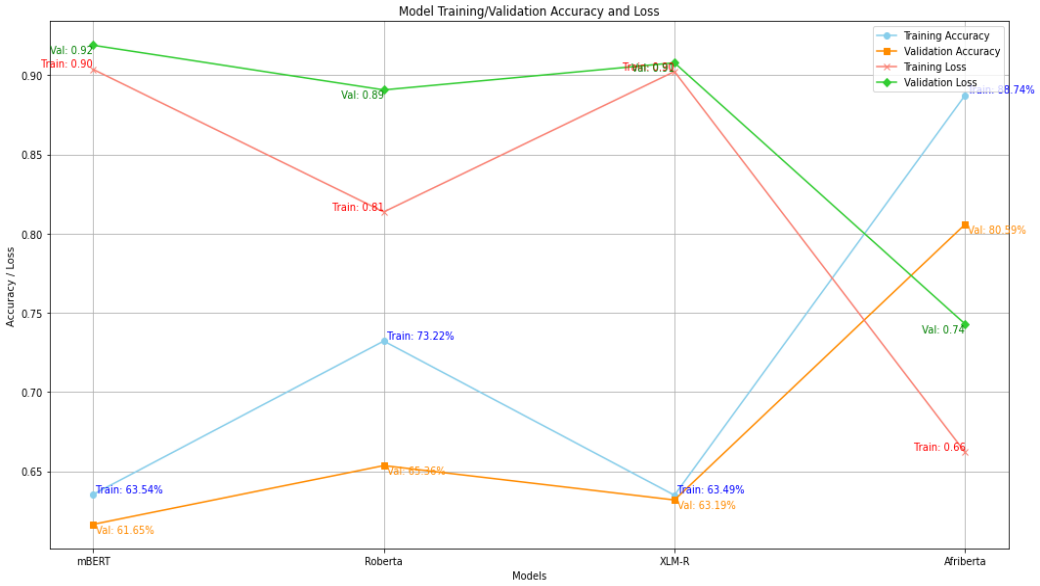


Fig. 7. Training / validation accuracy and loss (learning curve).

Table VIII provides a comprehensive comparison of the performance metrics benchmarking. It further illustrates the superiority of the proposed framework. These results signify the effectiveness framework. It demonstrated improvements over existing benchmarks underscore the novel contributions and practical utility of the framework through parameter tuning in low-resource language sentiment analysis.

TABLE VIII. PERFORMANCE BENCHMARKING COMPARISON

Reference	***F1-score***
[46]	75.1%.
[49]	71.2%.
[22]	78.3%
Ours	**80.6%**

VI. CONCLUSION AND FUTURE WORK

In conclusion, our comprehensive analysis of multilingual sentiment analysis models across a range of low-resource African languages provides valuable insights into their strengths and limitations. While AfriBERTa emerges as a robust performer, demonstrating effectiveness in capturing sentiment complexity the evaluation of mBERT, RoBERTa, and XLM-R reveals nuances in their training and validation metrics.

Despite the effective performance of the framework, a notable concern is the presence of overfitting. Most especially in on the AfriBerta suggests a need for addressing overfitting issues. To mitigate this, future work will incorporate data augmentation techniques during training, striving to improve the model's generalization to new data and enhance its performance on validation sets.

Furthermore, the study aims to expand its language coverage by experimenting with additional languages. While our current framework focuses on a diverse set of African languages, the inclusion of more languages will contribute to a more comprehensive understanding of the model's adaptability across different linguistic contexts. This expansion will involve fine-tuning the models on datasets specific to the additional languages, ensuring broader applicability of the sentiment analysis framework.

It is essential to acknowledge certain limitations in our study. One notable challenge is the prevalence of code-mixing in the tweets, where phrases of the English language are inserted within a single sentence or tweet. This linguistic phenomenon can impact the model's performance, and as such, future work will delve into addressing code-mixing challenges. Strategies such as incorporating code-switching-aware models or developing methods to handle mixed-language expressions will be explored to enhance the framework's efficacy.

The study provides valuable insights, but inherent challenges such as varying data quality across languages and potential overfitting in the training data must be acknowledged. Future research should address these limitations, explore additional low-resource languages, and enhance model interpretability. In conclusion, our comparative analysis adds to the expanding body of knowledge on sentiment analysis in low-resource languages, emphasizing the importance of linguistic diversity for model development in the African context.

ACKNOWLEDGMENT

We express our sincere appreciation to PTDF (Petroleum Technology Development Fund for their invaluable support in financing the scholarship. Additionally, we would like to appreciate the support received from the Universiti Teknologi PETRONAS (UTP) research grant: YUTP-FRG (015L0-312). Their involvement, financial assistance, and contributions to knowledge have played a pivotal role in facilitating the dissemination of our research findings, fostering engagement with emerging developments, and establishing meaningful professional connections within our field. We deeply value the investment made in our academic and practical development.

REFERENCES

[1] O. Alharbi, "A Deep Learning Approach Combining CNN and Bi-LSTM with SVM Classifier for Arabic Sentiment Analysis," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120618.

[2] R. Liu, Y. Shi, C. Ji, and M. Jia, "A Survey of Sentiment Analysis Based on Transfer Learning," IEEE Access, vol. 7. 2019. doi: 10.1109/ACCESS.2019.2925059.

[3] A. Altaf et al., "Deep Learning Based Cross Domain Sentiment Classification for Urdu Language," IEEE Access, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3208164.

[4] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," IEEE Transactions on Affective Computing, vol. 13, no. 2. 2022. doi: 10.1109/TAFFC.2020.2970399.

[5] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research," IEEE Trans Affect Comput, vol. 14, no. 1, 2023, doi: 10.1109/TAFFC.2020.3038167.

[6] B. Liu, "Sentiment analysis and subjectivity," in Handbook of Natural Language Processing, Second Edition, 2010.

[7] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowl Based Syst, vol. 226, 2021, doi: 10.1016/j.knosys.2021.107134.

[8] M. A. Paredes-Valverde, R. Colomo-Palacios, M. D. P. Salas-Zárate, and R. Valencia-García, "Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach," Sci Program, vol. 2017, 2017, doi: 10.1155/2017/1329281.

[9] A. Kumar and G. Garg, "Systematic literature review on context-based sentiment analysis in social multimedia," Multimed Tools Appl, vol. 79, no. 21–22, 2020, doi: 10.1007/s11042-019-7346-5.

[10] G. I. Ahmad, J. Singla, A. Ali, A. A. Reshi, and A. A. Salameh, "Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus: A Comprehensive Review," International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, 2022, doi: 10.14569/IJACSA.2022.0130254.

[11] M. Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," Inf Sci (N Y), vol. 512, 2020, doi: 10.1016/j.ins.2019.10.031.

[12] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, and A. Gelbukh, "Threatening Language Detection and Target Identification in Urdu Tweets," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3112500.

[13] A. I. Abubakar, A. Roko, A. M. Bui, and I. Saidu, "An Enhanced Feature Acquisition for Sentiment Analysis of English and Hausa Tweets," International Journal of Advanced Computer Science and Applications, vol. 12, no. 9, 2021, doi: 10.14569/IJACSA.2021.0120913.

[14] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges," ArXiv, Jun. 2020.

[15] D. D. Londhe, A. Kumari, and M. Emmanuel, "Challenges in Multilingual and Mixed Script Sentiment Analysis," in 2021 6th International Conference for Convergence in Technology, I2CT 2021, 2021. doi: 10.1109/I2CT51068.2021.9418087.

[16] Laumann Felix, "Low-resource language: what does it mean?," NeuralSpace. Accessed: Apr. 20, 2024. [Online]. Available: https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5

[17] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," in NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 2021. doi: 10.18653/v1/2021.naacl-main.201.

[18] M. M. Agüero-Torales, J. I. Abreu Salas, and A. G. López-Herrera, "Deep learning and multilingual sentiment analysis on social media data: An overview," Appl Soft Comput, vol. 107, 2021, doi: 10.1016/j.asoc.2021.107373.

[19] K. Dashtipour et al., "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques," Cognit Comput, vol. 8, no. 4, 2016, doi: 10.1007/s12559-016-9415-7.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, 2014, doi: 10.1016/j.asej.2014.04.011.

[21] J. Z. Maitama, N. Idris, A. Abdi, L. Shuib, and R. Fauzi, "A systematic review on implicit and explicit aspect extraction in sentiment analysis," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3031217.

[22] S. H. Muhammad et al., "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis," in 2022 Language Resources and Evaluation Conference, LREC 2022, 2022.

[23] A. Toktarova et al., "Offensive Language Identification in Low Resource Languages using Bidirectional Long-Short-Term Memory Network," International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.0140687.

[24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[25] N. H. M. Et.al, "Sentiment Analysis of Code-Mixed Text: A Review," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 3, 2021, doi: 10.17762/turcomat.v12i3.1239.

[26] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Systems with Applications, vol. 223. 2023. doi: 10.1016/j.eswa.2023.119862.

[27] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif Intell Rev, vol. 55, no. 7, 2022, doi: 10.1007/s10462-022-10144-1.

[28] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," Decision Analytics Journal, vol. 3, 2022, doi: 10.1016/j.dajour.2022.100073.

[29] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," Artif Intell Rev, vol. 54, no. 7, 2021, doi: 10.1007/s10462-021-09973-3.

[30] K. Cortis and B. Davis, "Over a decade of social opinion mining: a systematic review," Artif Intell Rev, vol. 54, no. 7, 2021, doi: 10.1007/s10462-021-10030-2.

[31] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics (Switzerland), vol. 9, no. 3, 2020, doi: 10.3390/electronics9030483.

[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, and ..., "Huggingface's transformers: State-of-the-art natural language processing," 2019.

[33] L. L. Maceda, A. A. Satuito, and M. B. Abisado, "Sentiment Analysis of Code-mixed Social Media Data on Philippine UAQTE using Fine-tuned mBERT Model," International Journal of Advanced Computer Science and Applications, vol. 14, no. 7, 2023, doi: 10.14569/IJACSA.2023.0140777.

[34] L. W. Astuti, Y. Sari, and S. -, "Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data," International Journal of Advanced Computer Science and Applications, vol. 14, no. 10, 2023, doi: 10.14569/IJACSA.2023.0141053.

[35] K. Subramanyam Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus: Una encuesta de modelos preentrenados basados en transformadores en el procesamiento del lenguaje natural," arXiv preprint arXiv ..., 2021.

[36] E. Ogbuju and M. Onyesolu, "Development of a General Purpose Sentiment Lexicon for {I}gbo Language," Proceedings of the 2019 Workshop on Widening NLP, 2019.

[37] L. Medrouk and A. Pappa, "Deep learning model for sentiment analysis in multi-lingual corpus," in Lecture Notes in Computer Science

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017. doi: 10.1007/978-3-319-70087-8_22.

[38] O. Rakhmanov and T. Schlippe, "Sentiment Analysis for Hausa: Classifying Students' Comments," in 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022.

[39] A. Ghafoor et al., "The Impact of Translating Resource-Rich Datasets to Low-Resource Languages through Multi-Lingual Text Processing," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3110285.

[40] M. Attia, Y. Samih, A. Elkahky, and L. Kallmeyer, "Multilingual multi-class sentiment classification using convolutional neural networks," in LREC 2018 - 11th International Conference on Language Resources and Evaluation, 2019.

[41] T. Isbister, F. Carlsson, and M. Sahlgren, "Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?," Apr. 2021.

[42] Z. Li, X. Li, J. Sheng, and W. Slamu, "AgglutiFiT: Efficient Low-Resource Agglutinative Language Model Fine-Tuning," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3015854.

[43] A. Jamatia, S. D. Swamy, B. Gambäck, A. Das, and S. Debbarma, "Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus," International Journal on Artificial Intelligence Tools, vol. 29, no. 5, 2020, doi: 10.1142/S0218213020500141.

[44] J. T. H. 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, 2023, doi: 10.3390/bdcc7020061.

[45] A. Yusuf, A. Sarlan, K. U. Danyaro, and A. S. B. A. Rahman, "Fine-tuning Multilingual Transformers for Hausa-English Sentiment Analysis," in 2023 13th International Conference on Information Technology in Asia (CITA), IEEE, Aug. 2023, pp. 13–18. doi: 10.1109/CITA58204.2023.10262742.

[46] S. H. Muhammad et al., "SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval)," in 17th International Workshop on Semantic Evaluation, SemEval 2023 - Proceedings of the Workshop, 2023. doi: 10.18653/v1/2023.semeval-1.315.

[47] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.747.

[48] J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, "Multilingual Language Model Adaptive Fine-Tuning: A Study on African Languages," in Proceedings of COLING 2022, 2022.

[49] S. Muhammad et al., "AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2023, pp. 13968–13981. doi: 10.18653/v1/2023.emnlp-main.862.