big data and cognitive computing

MDPI

Article

Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages

Koena Ronny Mabokela [1,*,†], Mpho Primus [2] and Turgay Celik [3,†]

[1] Department of Applied Information Systems, University of Johannesburg, Bunting Road Campus, Auckland Park 2092, South Africa
[2] Institute for Intelligent Systems, University of Johannesburg, Kingsway Campus, Auckland Park 2092, South Africa; mraborife@uj.ac.za
[3] Department of ICT, Centre for Artificial Intelligence Research (CAIR), University of Agder, 4879 Grimstad, Norway; turgay.celik@wits.ac.za
* Correspondence: rmabokela@uj.ac.za; Tel.: +27-79-711-3298
† Current address: School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2000, South Africa.

Abstract: Sentiment analysis is a crucial tool for measuring public opinion and understanding human communication across digital social media platforms. However, due to linguistic complexities and limited data or computational resources, it is under-represented in many African languages. While state-of-the-art Afrocentric pre-trained language models (PLMs) have been developed for various natural language processing (NLP) tasks, their applications in eXplainable Artificial Intelligence (XAI) remain largely unexplored. In this study, we propose a novel approach that combines Afrocentric PLMs with XAI techniques for sentiment analysis. We demonstrate the effectiveness of incorporating attention mechanisms and visualization techniques in improving the transparency, trustworthiness, and decision-making capabilities of transformer-based models when making sentiment predictions. To validate our approach, we employ the SAfriSenti corpus, a multilingual sentiment dataset for South African under-resourced languages, and perform a series of sentiment analysis experiments. These experiments enable comprehensive evaluations, comparing the performance of Afrocentric models against mainstream PLMs. Our results show that the Afro-XLMR model outperforms all other models, achieving an average F1-score of 71.04% across five tested languages, and the lowest error rate among the evaluated models. Additionally, we enhance the interpretability and explainability of the Afro-XLMR model using Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP). These XAI techniques ensure that sentiment predictions are not only accurate and interpretable but also understandable, fostering trust and reliability in AI-driven NLP technologies, particularly in the context of African languages.

Keywords: explainable AI; sentiment analysis; African languages; Afrocentric models; pre-trained models; transformer models

Citation: Mabokela, K.R.; Primus, M.; Celik, T. Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages. *Big Data Cogn. Comput.* **2024**, *8*, 160. https://doi.org/10.3390/bdcc8110160

Academic Editor: Jun Wang

Received: 27 August 2024
Revised: 27 October 2024
Accepted: 7 November 2024
Published: 15 November 2024

1. Introduction

Artificial intelligence (AI) has become increasingly popular, and, if used wisely and effectively, it has the potential to surpass our most optimistic expectations in a variety of practical applications. AI is seen as a facilitator in achieving the United Nations' Sustainable Development Goals (SDGs) by automating economic sectors [1,2]. A major challenge in AI development is the lack of explainability in many powerful techniques, especially those that have emerged recently. This includes large language models (LLMs) based on transformer models, ensemble methods, and deep neural networks (DNNs) [3,4]. XAI is an emerging field that focuses on introducing transparency and interpretability to complex DNNs or transformer-based LLMs. These models are often referred to as black-box models due to their inability to provide insight into their decision-making processes or

to offer explanations for their predictions and biases [5]. As illustrated in Figure 1, XAI refers to a collection of processes and methods that empower human users to understand and trust the outputs generated by machine learning algorithms [5,6]. This is essential to ensure transparent decision-making in AI applications and align with the SDGs [7,8]. XAI has demonstrated its value in medical imaging research [9], human–computer interactions, stock price prediction [10], and NLP tasks, providing insights into machine learning models' application processes and human-centered approaches [6,11,12]. Despite these advancements, a significant gap remains in integrating human factors into AI-generated explanations to make them more understandable and enhance performance in NLP for 41 African languages. XAI enables users to develop trust in NLP applications such as sentiment analysis, creating a positive feedback loop that continually improves the NLP model's performance.

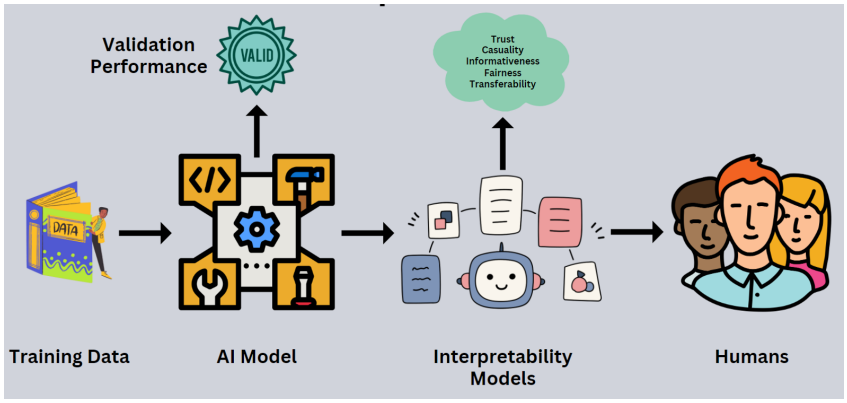


Figure 1. A step-by-step general-purpose framework for XAI systems.

Sentiment analysis is a crucial task in NLP that involves determining the emotional tone or attitude expressed in a piece of textual content [13]. It holds immense importance in numerous domains, such as marketing, customer service, and public opinion analysis on social challenges [14]. By automatically classifying text as positive, negative, or neutral, sentiment analysis provides valuable insights into customer feedback and helps businesses make informed decisions [13,14]. Recently, sentiment analysis research has expanded its focus from high-resourced languages to low-resourced languages, including over 2000 African languages [15]. Many researchers have concentrated on creating datasets and transformer-based NLP models to tackle the difficulties in African languages [16–19]. However, the interpretability and explainability of these models have not been explored. These languages are classified as low-resourced since they often lack sufficient labeled data and resources, thus presenting unique challenges for existing sentiment analysis. models. Moreover, due to the increased use of social media, particularly Twitter (now X), the use of multiple languages in multilingual communities is seen as a common practice, which further presents challenges in current models [19]. To address these issues, researchers have proposed various methods, such as transfer learning and multilingual and cross-lingual approaches [14], leveraging machine learning and DNN techniques [20]. Lately, we have witnessed the development of sentiment datasets for low-resourced languages, together with multilingual pre-trained language models (PLMs). These models are fine-tuned for specific NLP tasks, contributing to the improvement of these less privileged languages.

Established PLMs like BERT [21], RoBERTa [22], and XLM-R [23] have achieved impressive results (including state-of-the-art performance in some cases) on various NLP tasks, including sentiment analysis. Despite these PLMs being pre-trained in over 100+ languages, many African languages are still not well represented in the model pre-training process. Inspired by the success of mainstream PLMs, researchers have developed similar models specifically for African languages. Examples include AfriBERTa [24], Afro-XLMR [17], and AfroLM [18]. These Afrocentric PLMs are trained on massive datasets of African text data and can handle multiple African languages, promoting NLP advancements within

the continent. AfriBERTa, a pioneering African language model for transfer learning and fine-tuning in various NLP tasks, does not currently include South African languages [24]. This transfer learning approach allows us to adapt the model's knowledge to our target languages and improve its performance for sentiment analysis. Furthermore, a recent development is SERENGETI [16], a massively multilingual language model designed to cover a remarkable 517 African languages. This model holds promise for our sentiment analysis tasks, and we will investigate its suitability for our specific language set. Unlike mainstream PLMs, which often struggle with intricate tonal systems, complex morphology, and code-switching in African languages, Afrocentric models demonstrate promise for accurate sentiment analysis. However, a significant gap exists in understanding the internal workings of these models, specifically, how they represent and utilize features to classify text into sentiment categories. This lack of transparency hinders our ability to trust and interpret their decisions. Despite the growing advancements in NLP for high-resourced languages, sentiment analysis for low-resourced African languages remains largely underexplored. The scarcity of large, annotated datasets and the linguistic diversity of African languages pose significant challenges for developing robust sentiment analysis models. Furthermore, explainability in AI has not been extensively studied for these languages, limiting the trust and transparency of AI systems used in African contexts. This study was motivated by the need to bridge the digital divide in AI technologies for African languages. By developing explainable AI models for sentiment analysis, we aimed to improve transparency, trust, and adoption of AI systems in the African context. Our integration of Afrocentric PLMs with explainability techniques, such as LIME and SHAP, was intended to provide both effective sentiment analysis and insights into model decision-making, thereby fostering user confidence in AI applications.

Fortunately, XAI methods offer a solution to uncover how the black-box model makes a decision [25]. Using XAI techniques, we can better understand how multilingual PLMs work and how they figure out sentiment predictions. This improved understanding fosters trust and allows for targeted improvements in future model development for African languages [12]. To achieve this understanding, we applied XAI techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP), including their adaptation for transformer-based models. These techniques provide valuable insight into how Afrocentric PLMs make sentiment predictions. For instance, LIME can highlight specific words or phrases that significantly influence the model's sentiment classification. Similarly, SHAP can explain the contribution of each feature (word, n-gram, etc.) to the final sentiment score, allowing us to understand the reasoning behind the model's decisions [3]. We believe that this is the first investigation into XAI for sentiment analysis using Afrocentric PLMs adapted specifically for low-resourced South African languages. Additionally, by incorporating XAI techniques, we aim to unlock a deeper understanding of these models' decision-making for sentiment classification in the under-represented language group.

The significance of this study lies in its integration of Afrocentric pre-trained language models with XAI techniques (LIME, SHAP) specifically for low-resourced African languages. The study is the first to explore explainability in multilingual contexts, leveraging local linguistic features for improved sentiment analysis. Additionally, it demonstrates how these explainable models foster trust and interpretability in real-world applications across South African languages, such as Sepedi, Setswana, Sesotho, isiZulu, and isiXhosa. These languages are widely spoken in South Africa and extend to neighboring countries such as Botswana, Lesotho, Eswatini (formerly Swaziland), Mozambique, and Zimbabwe. SAfriSenti corpus is crucial for fine-tuning our models and evaluating their performance on sentiment analysis tasks. The use of the SAfriSenti corpus is crucial for fine-tuning our models and evaluating their performance on sentiment analysis tasks.

The main contributions of our study are summarized as follows:

- Our study is the first to integrate Afrocentric pre-trained language models (such as Afro-XLMR) with XAI techniques (LIME and SHAP) for sentiment analysis in

low-resourced African languages. While these models have been developed, their application in explainable sentiment analysis is novel and has not been explored extensively, especially in the context of African languages.
- We focus on South African languages (Sepedi, Setswana, Sesotho, isiZulu, and isiXhosa), which are significantly under-represented in the current NLP landscape. The study contributes to the field by addressing the unique linguistic challenges posed by these languages, and the results demonstrate how explainability can enhance AI models for these languages. This combination of explainability with Afrocentric models has not been explored before.
- We show how XAI methods like LIME and SHAP can provide insights into the decision-making process of models trained on multilingual datasets. This contributes to the growing field of interpretable AI for languages that lack linguistic resources.
- Our study also demonstrates real-world applicability by including human-centered evaluation of the explainability methods through participant feedback. This aspect of involving users to assess trust and transparency is unique, as it brings the study closer to practical use in African communities.

This study is structured as follows: In the next section, we provide a thorough literature review, discussing relevant research and recent developments in XAI methods for sentiment analysis, with a particular emphasis on explainability in transformer-based PLMs. Section 3 details the SAfriSenti corpus, the sentiment datasets used in our experiments. In Section 4, we explore the research conducted on XAI techniques and highlight the adaptation of Afrocentric PLMs for sentiment analysis. Section 5 outlines the experimental setup of the study, followed by the presentation and discussion of results. The limitations of the study are addressed in Section 6, and finally, Section 7 concludes the work, suggesting potential areas for future research.

2. Literature Review

In this literature review, we present the related work in the areas of XAI methods, XAI for sentiment analysis, and PLMs for sentiment analysis.

2.1. Sentiment Analysis Approaches

Sentiment analysis is a valuable tool for understanding public opinions using machine learning and NLP models [13,20]. Sentiment analysis has seen rapid growth and has been extensively explored across various domains, including text-based and multimodal sentiment analysis using diverse data formats such as text, audio, and video [13,26]. Several sentiment analysis approaches have been studied [13,27]. Ref. [28] employed capsule networks in conjunction with dependency-based concept-level analysis to effectively capture contextual relationships between words, resulting in enhanced sentiment classification performance. This study shares a similar goal of improving sentiment prediction accuracy, but we also focus on the interpretability of the predictions in low-resourced African languages. The study by [29] developed a framework that utilized various classifiers, including Bayes Net, KNN, C4.5 Decision tree, Support Vector Machine (SVM), and SVM with Particle Swarm Optimization (SVM-PSO), to extract opinions from comments in telemedicine videos. The SVM-PSO classifier demonstrated superior performance in assessing medical video reviews, achieving an accuracy of over 80% and outperforming other classifiers.

Recent research on sentiment analysis for low-resourced languages highlights the growing interest in developing effective approaches for these understudied languages. Transfer learning, word embedding, and machine translation are commonly used techniques [27]. Deep learning frameworks, particularly pre-trained transformers, have shown significant performance improvements [30]. Combining neural network architectures like Convolution Neural Networks (CNNs) and Bidirectional Long Short-term Memory (Bi-LSTM) with clustering techniques such as Gaussian mixture models (GMMs) and SVMs has demonstrated promising results for African languages [31]. However, challenges persist, including the scarcity of annotated datasets and standardized corpora [27,31]. Social

media platforms and product reviews are common data sources [30]. Various approaches, including machine learning, lexicon-based, and hybrid methods, have been proposed for multilingual sentiment analysis [32]. However, recent advances have leveraged deep learning architectures, such as Recurrent Neural Networks (RNNs) and transformer-based models like BERT, RoBERTa and GPT-2 [21]. These models, while significantly improving performance across multiple languages, are primarily trained on high-resource languages, posing challenges for under-resourced languages.

Since the introduction of the BERT model [21], XLM-R, and RoBERTa [23], PLMs built on the transformer architecture dominate the development of downstream NLP tasks such as sentiment analysis [22]. Compared to models that use deep learning architectures, PLMs have shown exceptional performance, consistently achieving state-of-the-art (SOTA) results [23]. Even though these models have demonstrated remarkable performance on sentiment analysis in high-resourced languages, their effectiveness in low-resourced languages has been a topic of research [17]. This has sparked interest in adapting PLMs to these low-resourced languages. In the domain of Afrocentric PLMs, several models have been developed for sentiment analysis and other NLP tasks, including AfriBERTa [16–18], to mitigate the lack of African languages in mainstream PLMs [22]. Ref. [24] developed the AfriBERTa model for 11 African languages. Ref. [17] presented an AfroXLMR pre-trained model for 17 African languages. Meanwhile, [18] investigated the AfroLM model for 23 African languages in NLP tasks. This current study introduces a novel approach that leverages the recently developed SERENGETI model [16], a powerful PLM designed specifically for 517 African languages. While recent advancements such as the Afro-XLMR and SERENGETI model [16] hold promise for broad applicability across African languages due to their massive language coverage, their effectiveness and the explainability of their predictions using XAI techniques remain unexplored.

2.2. Explainable AI Methods

The "explainable" aspect involves improving the transparency in how these models arrive at sentiment predictions [7]. This is crucial for understanding and validating model decisions, especially in critical applications such as social studies, market research, and policy analysis [33]. The human interpretation aspect of the explanation holds significant value. If the end-user of the application fails to understand it, its applicability to the end-user loses its purpose. Therefore, XAI methods are intended to assist domain experts in creating technically sound and human-interpretable explanations. They aim to make models intrinsically interpretable and transparent for black-box models [34]. To deal with black-box models, particularly transformer-based models, XAI methods were introduced to make the models interpretable.

XAI techniques like attention mechanisms and visualization are employed to provide insights into which parts of the input text contribute most to the sentiment analysis [12,35]. Ref. [36] introduced an attention-based explanation method for sentiment analysis using the BERT model. They demonstrated how attention scores reveal which parts of the input text contributed to the model's sentiment prediction. Ref. [37] extended the LIME framework to interpret the sentiment predictions of pre-trained models. Their approach highlighted keywords and phrases in the input text, making the predictions more transparent. Ref. [3] employed a TransSHAP that adapts SHAP to transformer models to understand BERT-based sentiment classifiers. They revealed the importance of individual words in the input text and provided insights into how the model makes its decisions. Ref. [38] proposed a novel approach that combines attention scores, integrated gradients, and LIME explanations to create a more comprehensive explanation framework for sentiment analysis using the BERT-like models. They used attention layers to extract explanation scores about model predictions and explore the internal behavior of the model.

Ref. [39] introduced LIME, a model-agnostic method for generating locally faithful explanations applied to sentiment analysis. Researchers have used LIME to perturb input text instances and build surrogate models that highlight the influential words affecting sen-

timent predictions [4]. LIME has proven effective in explaining the decisions of sentiment analysis models such as BERT and LSTM. Ref. [40] introduced SHAP values, a technique for explaining individual predictions in complex models. SHAP values have been applied to pre-trained models for sentiment analysis, quantifying the contribution of each word to sentiment prediction. Researchers have extended SHAP to consider both global and local explanations, providing a comprehensive understanding of model behavior [12,35]. Combining LIME and SHAP is a hybrid approach that leverages the strengths of both methods for interpreting sentiment predictions. This study emphasizes the value of cross-method validation to ensure robust explanations [7,41].

In [3]'s study, three visualization methods for explaining sentiment analysis were compared based on user feedback. LIME was favoured the most, receiving high ratings from 63.1% of respondents (with 39.1% giving it the top two scores). SHAP performed the least well, with an average score of 2.42 and 81.5% of the respondents giving it the lowest two scores [3]. When asked if they would use each method, 44.7% favored LIME, 42.1% favored TransSHAP, and only 34.2% favored SHAP. Overall, the results showed LIME as the most preferred method, with 54.3% of votes, followed by TransSHAP with 40.0%, while SHAP received the lowest preference at 5.7%. Hence, for this study, we explore a combination of these XAI methods to achieve the objectives of our study. Additionally, to ensure that model output aligns with human values, we incorporated a feedback process involving human input through a survey.

3. Datasets

Table 1 illustrates the statistical distribution of tweets across five South African languages. These datasets are categorized into two sections: The SAfriSenti corpus and the newly built dataset for two additional languages. These datasets were collected between September 2021 and May 2022. This dataset was collected before the Twitter/X platform was changed to its official name, X. In this study, we still refer to the textual information of the platform as tweets. In this study, we use ISO 639 codes for our languages. The ISO 639 standard provides codes for the representation of names of languages, including Sepedi, Setswana, Sesotho, isiXhosa, and isiZulu as illustrated in Table 1 [42].

Table 1. Distribution of positive, negative, and neutral sentiments across South African languages.

Language (ISO 639)	Positive		Negative		Neutral		Total
	Pos	%	Neg	%	Neu	%	
Sepedi (nso)	5153	48%	3270	30%	2355	22%	10,778
Setswana (tsn)	3932	51%	2150	28%	1590	21%	7672
Sesotho (sot)	3050	48%	2024	32%	1241	20%	6314
isiXhosa (xho)	6657	26%	12,125	48%	6421	26%	25,203
isiZulu (zul)	19,252	43%	22,400	49%	3378	8%	45,303

First, we describe the collection of texts used for training and fine-tuning multilingual transformer-based models developed for sentiment classification tasks. Then, we provide information about the sentiment dataset used to address the different cases of explainable sentiment analysis. These datasets are the gold standard with labels of three classes (positive, negative, and neutral), as shown in Figure 2. These sentiment datasets are further described below:

- SAfriSenti Corpus—a Twitter-based sentiment corpus developed for South African low-resourced languages in a multilingual context [19]. It is to date the largest sentiment corpus (i.e., over 40,000 tweets) for South African languages such as Sepedi, Setswana, and Sesotho, including English. The SAfriSenti corpus was manually annotated by experts and native speakers. Annotators were carefully selected based on their demonstrated proficiency in the native language and relevant technical skills required for the annotation process. Each tweet in the corpus underwent data cleaning and preprocessing steps where noise, URLs, and meaningless tweets were removed [43].

We replaced all @mentions by @user for data protection purposes [44]. Furthermore, the corpus consists of 64% monolingual tweets and 36% multilingual tweets, including code-switched tweets. We used Krippendorff's Alpha method to obtain an acceptable annotator reliability score.

- Additional Dataset—We used the Twitter/X API to collect this dataset, following the approach presented in [43]. This sentiment dataset was collected from Twitter and contains over 50,000 tweets in the isiZulu and isiXhosa languages. To automate sentiment labeling for our dataset, we leveraged a semi-automatic distant supervision approach. This method utilizes sentiment-bearing emojis and word-based sentiment lexicons, as detailed in [19]. However, we employed three native speakers to manually double-check and annotate only less than a 24% portion of the dataset in each language, as proven in [19]. In this corpus, the isiXhosa comprises 35.74% of the total tweets. Meanwhile, the isiZulu contains 64.25% of the total tweets. The dataset also includes code-switched English words.

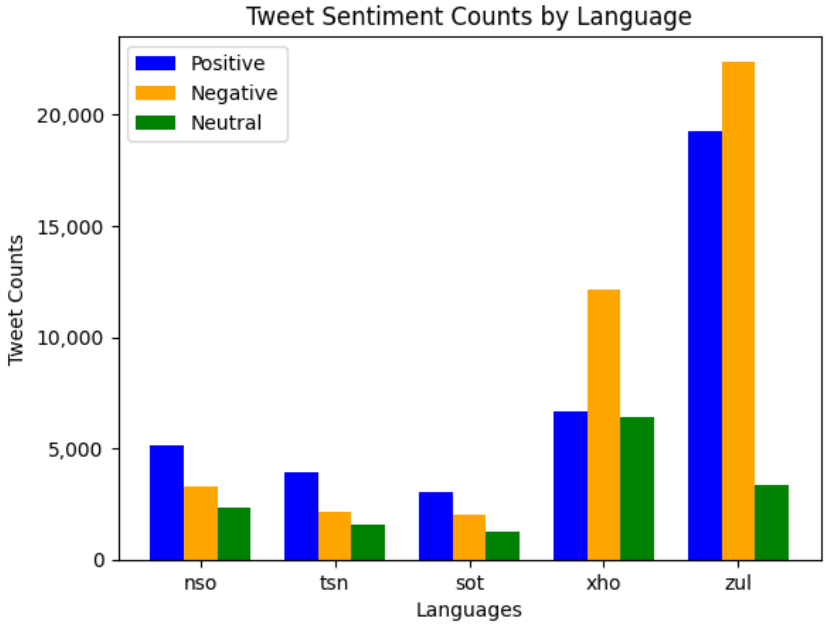


Figure 2. Distribution of tweets across languages. We show the graphical comparison of the various tweets across the all languages for sentiment analysis.

4. Proposed Methods for Explainable Sentiment Analysis

In this section, we describe our proposed approaches for this study. Figure 3 illustrates the workflow of the proposed approach for sentiment analysis using XAI techniques and fine-tuned models for low-resource African languages, specifically leveraging the SAfriSenti corpus. The process begins with a pre-trained language model that has been trained on a large-scale text corpus. This provides the initial general knowledge that the model has learned from vast amounts of text data. The pre-trained model is fine-tuned on the SAfriSenti corpus. This fine-tuning process produces the best model for sentiment analysis on African language datasets. The fine-tuned model is then used to generate sentiment predictions for the input text. Alongside this, XAI techniques are employed to provide transparency into how the model made its predictions. The goal of XAI is to make the decision-making process interpretable and understandable. The XAI outputs (i.e., explanations for the sentiment predictions) are presented through a visual interface. These explanations allow for human-centered decision-making, where users can interpret the model's reasoning and provide feedback. This interaction improves user trust and facilitates a better understanding of how AI models operate, particularly in African language contexts.

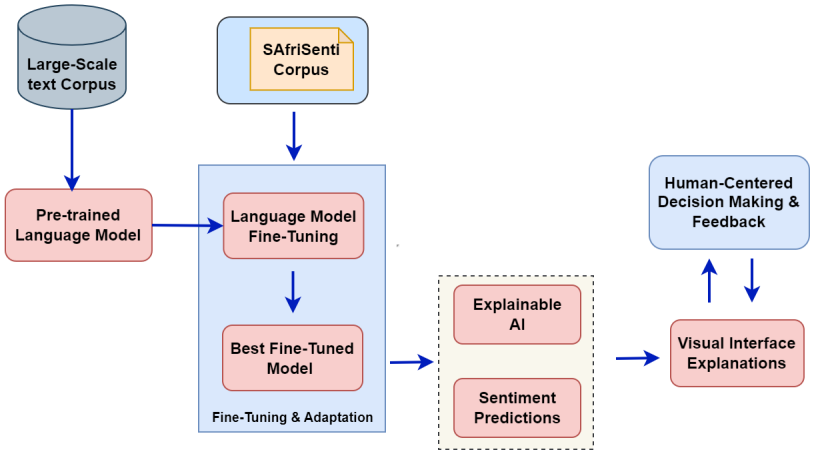


Figure 3. Overview of explainable AI framework for sentiment analysis.

4.1. Fine-Tuning of the PLMs

Figure 3 provides an overview of the system framework dedicated to leveraging PLMs for XAI methods in sentiment analysis. Fine-tuning PLMs for a specific downstream task, such as sentiment classification, is a common approach in NLP research [21]. To achieve this, we used a labeled portion of our sentiment dataset. An equation presented for fine-tuning a pre-trained model was considered:

$$\theta^* = arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_M(x_i'; \theta), y_i)$$

where:

- θ: The parameters of the model that are updated during training.
- θ^*: The optimal parameters found after minimizing the loss function.
- n: The number of training examples in the dataset.
- $\mathcal{L}$: The loss function, which measures the difference between the predicted and true labels.
- $f_M(x_i'; \theta)$: The output of the model M when applied to the input x_i', parameterized by θ.
- x_i': The preprocessed input text (e.g., tokenized tweet).
- y_i: The true label corresponding to the input x_i'.

Thereafter, we use XAI techniques to provide a clear and understandable explanation of how Africa-centric models achieve sentiment predictions in low-resourced African languages.

4.2. Explainability Methods

Explanation can be global (structure-based) or local (prediction-based). In this study, our objective is to help end-users understand the individual sentiment predictions made by Afrocentric PLMs [45]. Hence, we utilize XAI methods for local explainability, which specialize in explaining individual predictions. Furthermore, this research work demonstrates how the model pays attention to various word tokens or phrases in the tweets by using explainability methods that reveal attention visualization [34]. We need the weights for the encoder–decoder attention layers to calculate the attention per token. However, attention serves a purpose beyond just optimization; it also plays an essential role in expanding the context of the transformer-based language model [46]. We use LIME (https://github.com/shap/shap (accessed on 17 October 2024)) and TransSHAP (https://github.com/enjakokalj/TransSHAP (accessed on 17 October 2024)) (i.e., a version for transformer-based models) together with tools such as Bertviz to generate visualizations for further explanations [40]. The most widely used perturbation-based explanation

methods are LIME with a wrapper function to support the transformer-based model, and SHAP [5,41]. For SHAP to work effectively with BERT-like models, particularly Kernel SHAP, a function classifier that returns probabilities is used [3]. However, since the model-agnostic Kernel SHAP method does not support BERT-like models, we address this limitation by following the approach of developing a custom function to preprocess the input data and obtain predictions from the PLM model [3]. Thereafter, we prepare data for visualization within the TransSHAP framework based on the game theory optimal Shapley values [47]. The SHAP value of a particular feature for a data point is described by Equations (1) and (2):

$$\Phi_i = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{i\}} \frac{|S|!(p-|S|-1)!}{p!} [f_x(S \cup \{i\}) - f_x(S)] \quad (1)$$

$$f_x(S) = \mathbb{E}[f(x)|x_S] \quad (2)$$

- $f_x(S)$ is the expected model prediction given the subset of features S.
- $\mathbb{E}[f(x)|x_S]$ is the expected value of the model's prediction conditioned on the feature subset S.
- x is the vector of feature values of a tweet (which needs to be explained).
- S denotes a subset of input features, and Shapley values can be obtained through a value function (f_x).
- p is the number of features.
- $E[f(x)|x_S]$ is the expected value of the function on subset S.

The LIME method makes individual tweet classification predictions more understandable [39]. LIME achieves this by generating perturbed versions of the original tweet and evaluating their impact on the model's prediction [47]. The process of creating local surrogate models with interpretability constraints can be expressed as follows (Equation (3)):

$$Explanation(x) = arg \min_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (3)$$

- $Explanation(x)$: the LIME explanation for the model's prediction on input x (e.g., a tweet).
- $g \in G$: a simple interpretable model that approximates the original model f locally around the input x.
- $L(f, g, \pi_x)$: a loss function that measures how well the simple model g approximates the original model f near x. π_x is a proximity measure that weights the importance of different points in the neighborhood of x.
- $\Omega(g)$: a complexity penalty that discourages overly complex explanations (e.g., the use of too many features).

An interpretable model f can be considered for a tweet sample x that decreases the loss $(\mathcal{L})$ and computes how close the explanation is to the predictions provided by the model f. The complexity of the model $\Omega(g)$ is therefore maintained at a minimum level. The collection of realizable interpretations is denoted by $\mathcal{G}$. The closeness measure π_x expresses the extent of the locality around the tweet sample x.

In this study, we used both LIME and TransSHAP visualizations, since they are useful in sentiment explanations. However, we take a different research direction in evaluating the trust, transparency, reliability, and interpretability of the explanations provided by LIME and TransSHAP.

4.3. Pre-Trained Models Description

Fine-tuning involves further training of the models in our labeled dataset. This process updates the model parameters and weights to suit the downstream task. Furthermore, the model learns from the given labeled dataset and adjusts its parameters to make sentiment predictions. We conducted fine-tuning and adaptation processes on XLM-R, mBERT,

AfroLM, Afro-XLMR, and SERENGETI for the task of sentiment classification. Further elaborations on these models are presented below:

- mBERTis a multilingual version of BERT pre-trained in the top 104 languages with the largest Wikipedia dataset using a masked language modeling (MLM) objective and the next-sentence prediction (NSP) task [21]. We fine-tune the `multilingual-based` model with 172M model parameters by adding a linear classification layer on top of the pre-trained transformer model.
- XLM-R (XLM-RoBERTa) is a powerful multilingual model trained on a massive dataset of crawled text from hundreds of languages [23]. The XLM-R model is created by distilling knowledge from the DistilRoBERTa model into the XLM-R model using more than parallel data points from 50+ languages. The model was trained with the multilingual MLM objective [21]. XLM-R has 12 layers, a hidden size of 768, 12 attention heads, and a total of 270M parameters.
- AfroLM is a multilingual language model that has been pre-trained from scratch on African languages using a novel self-active learning framework [18]. Afro-XLMR is a state-of-the-art Transformer model pre-trained with 23 African languages, including Setswana, isiXhosa, and isiZulu. AfroLM uses a shared SentencePiece vocabulary with 250,000 byte pair encoders across the 23 languages. The model is trained primarily using the MLM objective. Trained on a considerably smaller dataset than existing models, it still surpasses many multilingual PLMs in various NLP tasks.
- Afro-XLMR was developed by adapting the XLM-R large model using the MLM objective on 17 African languages. This covers major language families such as Sesotho, isiXhosa, and isiZulu, along with three high-resource languages including English [17]. Afro-XLMR utalize multilingual adaptive fine-tuning that allows multilingual adaptation and preserves downstream performance on both high- and low-resourced languages. Afro-XLMR comes in different configurations such as base and large, with the large version having 278M parameters. We are motivated to use this model because the authors are confident that it can be easily adapted to a wide range of other African languages, including those with limited linguistic resources.
- SERENGETI is the largest African MPLM that was pre-trained using 42 GB of data comprising multiple domains from religious, news, government documents, health documents, and existing Wikipedia corpora [16]. The pre-training data cover 517 African languages and the 10 most spoken languages globally. This model was pre-trained on both Electra style [48] and XLM-R-style architecture. Electra utilizes the multilingual replaced token detection (MRTD) objective during training. The model uses 12 layers and 12 attention heads, with a sequence length of 512 and 277 M parameters. We used the version with a vocabulary size of 250K (SERENGETI-E250) WordPiece tokens. The SERENGETI model has significantly outperformed AfriBERTa, XLMR, mBERT, and Afro-XLMR on some NLP tasks.

5. Results and Discussion

This section outlines our experimental setup for fine-tuning models and evaluating sentiment analysis systems. We present and discuss the performance metrics, discuss attention mechanisms, and visualize the outcomes of XAI techniques.

5.1. Experimental Setup

We used the SAfriSenti corpus for evaluating our sentiment analysis models. The dataset for each of the five languages used in the experiments was divided into 80% for training and 20% for testing. The total number of training tweet samples is 36,023, with isiZulu having the largest portion at 15,496 tweet samples, followed by isiXhosa with 5281, Sepedi with 7168, Setswana with 2874, and Sesotho with 2289. The testing dataset consists of 9007 tweet samples, with isiZulu again having the largest share at 3756 tweet samples, followed by Xhosa with 1376, Sepedi with 1793, Setswana with 707, and Sesotho with 298 tweet samples. This split reflects a significant size difference between the training and

testing sets, ensuring a sufficient number of tweet samples for both training and evaluation. We performed model fine-tuning by considering the final hidden vectors of the first special token as the aggregate input sentence representation and then passed them onto the softmax classification layer to obtain the predictions. We fine-tuned our multilingual PLMs with the hyperparameters specified in Table 2. All training experiments were performed using the HuggingFace Transformers library. We conducted our experiments using the Google Colab Pro service, which provides access to high-performance GPUs. Specifically, we were assigned an NVIDIA Tesla T4 GPU with 15 GB memory. Google Colab Pro also offers enhanced computational resources, including increased CPU and RAM, which supported the efficiency of our experiments. We verified the GPU configuration at the beginning of each session to ensure consistency in computational resources across our experiments.

Table 2. Hyperparameters used for fine-tuning the PLMs.

Model	Hyper-Parameters	Values
Our PLMs	sequence maximum length	256
	hidden size	768
	attention heads	6
	hidden layers	10
	learning rate	1×10^{-4}
	batch size	16/32
	AdamW optimizer	-
	Epochs	10
	Dropout rate	0.2

In the fine-tuning process, we froze the transformer layers for models like AfroLM, SERENGETI and AfroXLMR. Only the task-specific sentiment classification layers were fine-tuned, allowing the model to leverage the pre-trained knowledge while reducing the risk of overfitting on the subsets with smaller datasets. Similarly, for models like mBERT and XLM-R, we froze the lower layers to maintain the generalization power of the pre-trained embeddings. To efficiently fine-tune a PLM with limited annotated data, freezing specific layers can substantially reduce the number of parameters, making the process more feasible in constrained environments [49]. To optimize the model performance, we performed hyperparameter tuning using a grid search over learning rates (ranging from 1×10^{-5} to 1×10^{-3}), batch sizes (16, 32), and dropout rates (0.1 to 0.5). The learning rate for each model was carefully selected based on a grid search, with batch sizes ranging from 16 to 32 depending on model size and dataset characteristics. Only XML-R models perform well with a batch size of 16 and learning rate of 2×10^{-4}.

Moreover, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority classes in each sentiment dataset [50]. This technique generates synthetic samples by interpolating existing samples, thus balancing the class distributions. Additionally, we employed class weighting during model fine-tuning, allowing the model to assign more importance to the minority classes and mitigate the risk of the model favoring the majority class. We utilized several strategies to prevent overfitting: (i) we applied k-fold cross-validation (k = 5), ensuring the model's generalization across different data partitions and reducing the likelihood of overfitting; (ii) we introduced early stopping by monitoring the validation loss and halting training when no further improvement was observed; and (iii) we incorporated dropout layers in models, which is a regularization technique used to prevent overfitting.

5.2. Evaluation Metrics

To evaluate the performance of the models, we used the F1-score. Given the class imbalance, the F1-score was the primary metric for assessing the model's performance, as it

balances precision and recall [20]. The F1-score is a metric used to evaluate the performance of machine learning models, particularly in classification tasks [30]. It summarizes a model's accuracy in a way that balances the trade-off between precision and recall. The F1-score can be calculated using the following formula:

- True Positives (TPs): the number of tweets correctly labeled as having a positive sentiment.
- True Negatives (TNs): the number of tweets correctly labeled as not having a positive sentiment.
- False Positives (FPs): the number of items incorrectly labeled as positive sentiment.
- False Negatives (FNs): the number of items incorrectly labeled as not having a positive sentiment.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

where precision is the ability of a model to avoid labeling negative instances as positive and recall is the ability of a model to find all positive instances. We reported the F1-score for each of our models, individually for each language.

5.3. Performance Results

Table 3 presents a comparative analysis of the F1-score performance of five language models (mBERT, XLM-R, AfroLM, Afro-XLMR, and SERENGETI) across five African languages: Sepedi (nso), Setswana (tso), Sesotho (sot), isiXhosa (xho), and isiZulu (zul). The F1-score is used as a balanced metric that considers both precision and recall and offers a reliable measure of the effectiveness of each model in sentiment analysis for these under-resourced languages.

Table 3. Language model F1-score performance comparison.

Language	mBERT	XLM-R	AfroLM	Afro-XLMR	SERENGETI
nso	57.69%	48.20%	76.69%	54.54%	51.75%
tso	61.54%	54.44%	64.23%	55.23%	58.01%
sot	65.21%	55.31%	58.32%	65.49%	61.20%
xho	85.78%	72.31%	75.32%	89.17%	80.01%
zul	86.46%	74.60%	82.77%	90.74%	79.39%

With the use of the languages isiZulu and isiXhosa, as well as Sepedi, Setswana, and Sesotho, which share linguistic similarities and characteristics, Afro-XLMR also performs extremely well, demonstrating its adaptability to multiple African languages. This is more likely due to fine-tuning language-specific data (in-domain dataset) and linguistic features. Also, this is due to Afro-XLMR being initially pre-trained in some of these languages (Sesotho, isiXhosa, and isiZulu). Furthermore, the superior performance of Afro-XLMR in isiXhosa and isiZulu can be attributed not only to its pre-training on these languages but also to its ability to better capture the syntactic and semantic properties specific to Bantu languages. Additionally, the mBERT model also performs better on isiZulu and isiXhosa languages that share linguistic similarities and characteristics. The pre-training phase of the mBERT model involved these two languages, which contributed to the notable performance improvement of over 80%. Furthermore, the highest quantity and quality of training data for each language also affect performance in the isiXhosa and isiZulu languages. Furthermore, future research and model fine-tuning may be necessary to improve sentiment analysis performance in some languages, especially those with

lower sentiment scores, such as Setswana and Sesotho. The relatively lower performance in Setswana and Sesotho suggests that these languages may require additional model fine-tuning. One possible solution is the use of transfer learning from better-performing languages within the same family, or implementing data augmentation techniques to increase the quantity and diversity of the training data.

Figure 4 compares the performance of five different language models (mBERT, XLM-R, AfroLM, Afro-XLMR, and SERENGETI) in five different languages (nso, tso, sot, xho, and zul). Performance is measured using the F1-score. Afro-XLMR generally performs the best in all languages, followed closely by XLM-R. mBERT and SERENGETI show comparable performance, while AfroLM tends to lag behind the others. XLM-R performs the best, followed closely by Afro-XLMR and mBERT in Sepedi. Afro-XLMR performs the best, with XLM-R and mBERT not far behind in Setswana. Afro-XLMR again takes the lead, followed by XLM-R and SERENGETI in Sesotho. In isiXhosa, Afro-XLMR and XLM-R are the top performers, with SERENGETI and AfroLM close behind. AfroLM performs best in isiZulu but still falls short of Afro-XLMR. XLM-R and SERENGETI also perform well. The results suggest that models specifically trained in African languages, such as Afro-XLMR, tend to perform better than more models trained on monolingual datasets. This highlights the importance of developing language models that are specific to under-represented languages. Our results are comparable to those of previous work [18,44]. The relatively strong performance of SERENGETI, AfroLM and Afro-XMLR across different languages suggests that it might be a good choice for general-purpose sentiment analysis tasks for African languages.

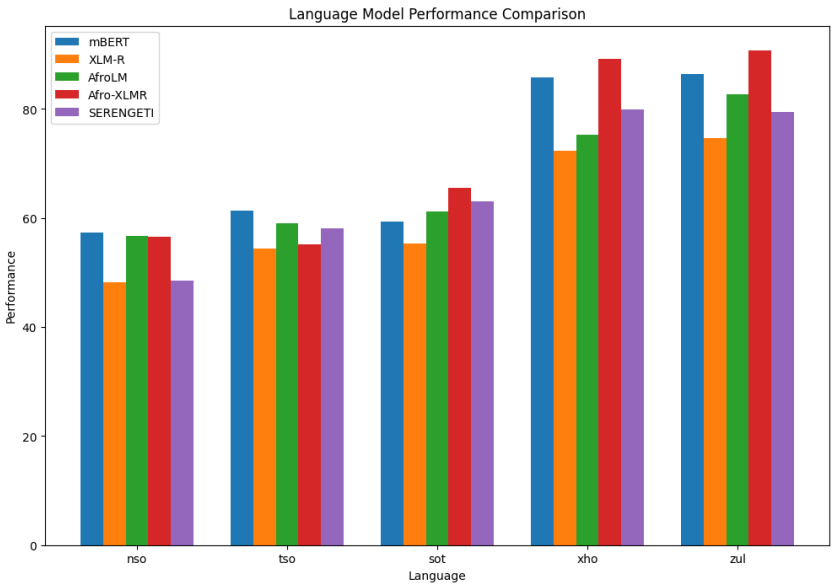


Figure 4. Language model performance comparison. The F1-score is measured as a percentage (%).

Our proposed Afro-XLMR model outperformed existing models like mBERT, XLM-R, and AfroLM on sentiment analysis across all five languages (Sepedi, Setswana, Sesotho, isiXhosa, and isiZulu). As shown in Figure 4 above, Afro-XLMR achieved the highest F1-scores in isiXhosa and isiZulu, surpassing mBERT by up to 15% and XLM-R by 7% for these languages. However, it is also noted that Afro-XLMR achieved an average F1-score of 71.04%, outperforming mBERT and XLM-R by 5% and 4%, respectively, across five tested languages. SERENGETI, another Afrocentric model, also demonstrated strong performances with an F1-score close to Afro-XLMR, especially in isiZulu. The superior performance of Afro-XLMR can be attributed to its fine-tuning of in-domain African language datasets. Compared to mainstream models like mBERT and XLM-R, Afrocentric models, when fine-tuned with the African language dataset, better adapt to the unique linguistic features prevalent in these languages. This highlights the need for tailored

language-specific models in low-resourced African language settings; it also highlights the effectiveness of our approach.

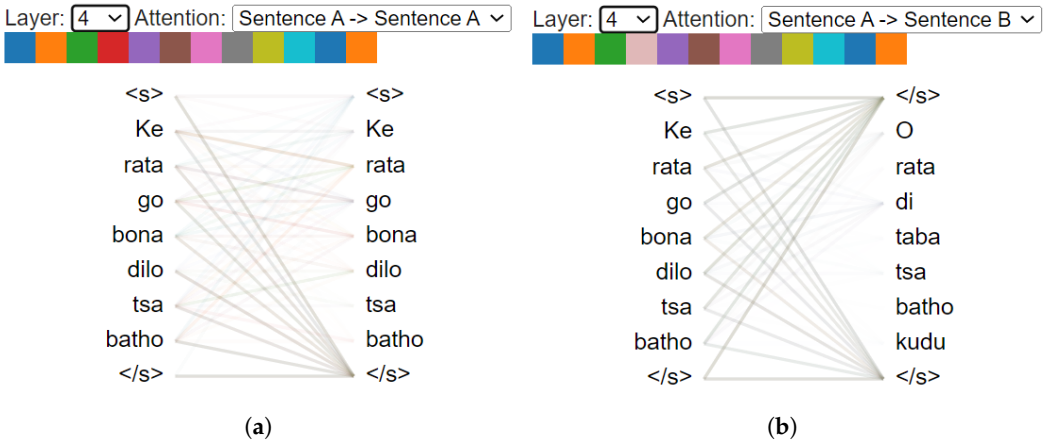
5.4. Attention Explanations

In this section, we visualize the attention of the model used for the explanations. Attention allows the model to weigh the importance of different words in a sentence when understanding the meaning of a particular word. It can be equated with the model focusing its "attention" on certain parts of the sentence. Let us consider the two sentences: sentence A: "Ke rata go bona dilo tsa batho"/I like to look at other people's belongings/and sentence B: "O rata di taba tsa batho kudu"/You like other people's business/affairs a lot/.

Figure 5a demonstrates "self-attention" where a sentence is compared to itself. This helps the model understand the internal relationships within the sentence. The visualization illustrates self-attention, where a sentence attends to itself. Each word in the sentence is assessing its relationship with every other word. The words of sentence A are vertically listed on both sides. Each line connects a word on the left with a word on the right. The opacity (thickness) of each line indicates the strength of the attention relationship between the connected words. The most opaque lines reveal the strongest attention relationships within the sentence. In this visualization, there is a high degree of self-attention. Each word seems to focus heavily on itself, indicating a strong sense of individual importance.

There are also notable connections between words like "bona" (see) and "dilo" (things), suggesting that the model recognizes the semantic relationship between these words. The faintest lines represent the weakest attention relationships. These words may be considered less relevant to each other in the context of the sentence. The strong pattern of self-attention indicates that the model has learned to emphasize individual words, possibly because of the importance of each word in conveying meaning in this specific sentence. The connections between words like "bona" and "dilo" show the model's ability to capture semantic relationships within the sentence, contributing to a richer understanding of the text.

Figure 5b shows the visualization of the attention mechanism in layer 4 of a transformer model, specifically illustrating the attention between two sentences (sentence A and sentence B). The words of both sentences are listed, A on the left and B on the right. Each line connects a word in A to one in B. Thicker lines mean stronger attention between the connected words. In this case, all lines seem equally faint. This suggests a uniform attention distribution, where each word in sentence A is paying roughly equal attention to all words in sentence B. This could indicate that the model is struggling to identify clear relationships between the two sentences in this layer.



(**a**) (**b**)

Figure 5. Comparison of attention visualizations. (**a**) Self-attention, (**b**) attention between two sentences. The intensity (brightness) of the color indicates the strength of the attention weight. Different colors are used to visually distinguish between different attention weights, making it easier to see the pattern of attention.

5.5. XAI Explanations

We generated all XAI visualizations using the Afro-XMLR model due to its superior performance. The visualization techniques employed in the explanation methods LIME and SHAP are primarily designed for interpreting tabular data. We used the LIME method to create visualizations, such as the one in Figure 6, to help us understand which words in parts of the tweets have the greatest influence on the model's final prediction. The model analyzed the text and determined a 57% probability of being negative and a 43% probability of being positive. From the example "Ke dula ke lapile and I am annoyed" (I am always tired and I am annoyed), we observe that the model successfully identifies the negative sentiment in the tweet.

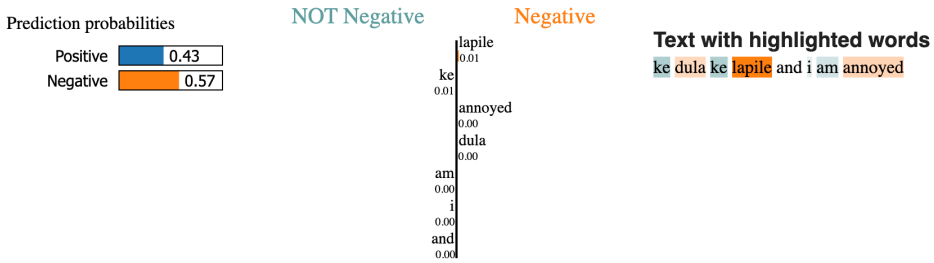


Figure 6. The LIME visualizations of a negative tweet in Sepedi mixed with English.

The model is paying attention to words that are negative in this context. The fact that words receiving the most attention can form a plausible explanation is due to the model's focus on relevant words, such as "lapile", "tired", and "annoyed" (orange), which indicate the negative sentiment of this tweet. Furthermore, the model also highlights words that are not negative, like "ke" and "I am" (blue). LIME focuses on the words that most influenced the model's decision. In this case, the words "lapile" (tired) and "ke" (I) are highlighted as the most influential for a negative prediction. The numbers next to each word (e.g., 0.01 for "lapile") represent the weight or importance assigned to that word by the simpler, local model in determining the sentiment. Here, the model seems to associate "lapile" and "ke" with negative sentiment, even though "ke" might not be inherently negative in all contexts.

The LIME results in Figure 7 show the prediction probabilities for the text "kodwa kukhulakala kuni lokhu ngempela" (but this really makes sense to you), the model predicts a negative sentiment with a probability of 0.63 and a positive sentiment with a probability of 0.37. The words contributing the most to the negative prediction are "kukhohlakala" and "ngempela", while "lokhu" and "kodwa" contribute the most to the positive prediction. The overall prediction is negative, suggesting that the model interprets the text as expressing a negative sentiment. Moreover, the LIME results in Figure 8 show the prediction probabilities for the text "mpilo bani ngempela lena" as negative (0.60) and positive (0.40). The word contributing the most to the negative prediction is "ngempela", while "mpilo" contributes the most to the positive prediction. The words "bani" and "lena" have a lower impact on the overall prediction. Despite contributions to the positive class, the overall prediction remains negative.

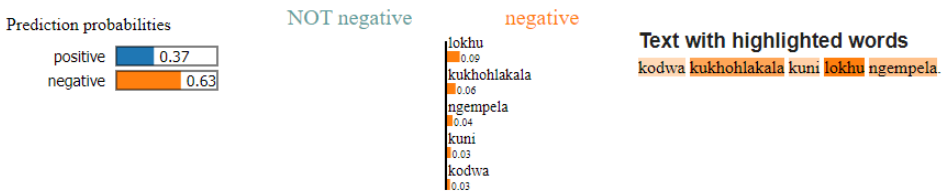


Figure 7. The LIME visualizations of a negative tweet in the isiZulu language.

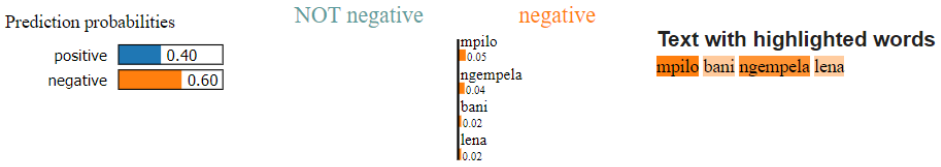


Figure 8. The LIME visualizations of the tweet in isiZulu with negative words highlighted.

The model incorrectly classified the tweet in Figure 9 as "negative", despite the original tweet expressing positive sentiment about going home. This discrepancy indicates a potential error in the model's understanding of the text. LIME highlights the words it believes contributed most to the prediction. Surprisingly, none of the words have a strong influence towards either positive or negative sentiment, as indicated by their low values close to 0. The word "rata" (like) has the highest value but it is still only 0.01. This suggests that the model might be struggling to interpret the overall meaning of the sentence due to context or nuances it does not fully grasp. The prediction probabilities are split evenly at 0.50 for both positive and negative, reflecting the model's uncertainty in classifying this particular tweet.

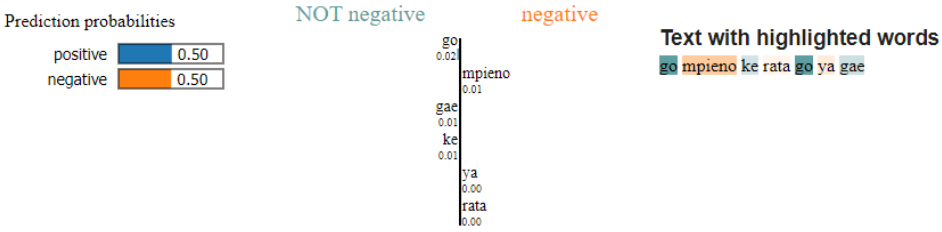


Figure 9. The LIME visualizations of an incorrect prediction in the Setswana language.

The SHAP visualization for the same sentence in Figure 10 shows that features with the strongest impact on the prediction correspond to longer arrows pointing towards the predicted class, which is negative. This suggests that a model with higher accuracy tends to generate more reliable explanations, as accurate predictions make the explanations more likely to be reliable and meaningful. The base value here leans slightly towards negative sentiment (0.697), representing the model's average prediction across the dataset. For the input text "ke dula ke lapile and I am annoyed", the output value is nearly 1.0, indicating a strong negative sentiment prediction. Each word in the input is a feature, and the plot shows the impact of each word on the model's prediction, moving it from the base value to the output value. Red bars indicate features that push the prediction towards negative sentiment, while blue bars push towards positive sentiment. The length of each bar represents the magnitude of its impact.

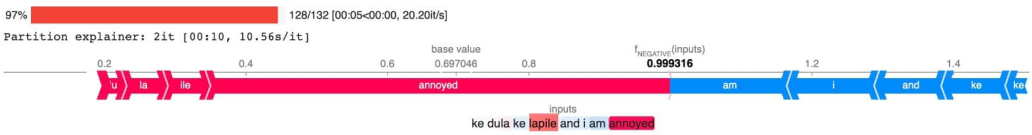


Figure 10. The SHAP visualizations of a tweet with a negative sentiment.

The SHAP results for the text "mpilo bani ngempela lena" in Figure 11 indicate a higher probability of negative sentiment (0.64) compared to positive (0.36). The words "lena" and "mpilo" contribute slightly towards the positive sentiment, while "ngempela" strongly pushes the prediction towards negative sentiment. Although there are some positive influences, the overall sentiment leans towards negative due to the significant impact of the word "ngempela." The base value of 0.5 represents the average prediction in the absence of any input features.

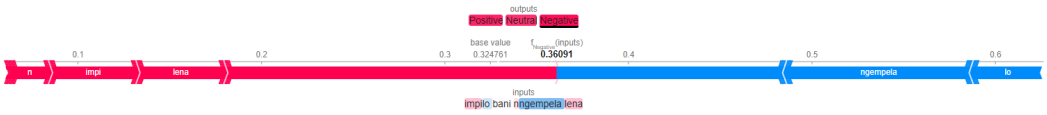


Figure 11. The SHAP visualizations of a tweet with negative sentiment.

In Figure 12, SHAP results for the Sepedi text "ke dula ke dumedisa le batho kudu" (I am always greeting people) indicate a high probability of a positive outcome with an output value of 0.6914. The base value of 0.34094 represents the average outcome across the entire dataset. The words "batho" and "kudu" contribute the most to this positive prediction, with "batho" having the highest impact. The other words in the sentence, such as "ke", "dula", "dumedisa", and "le," also have a positive impact but to a lesser extent. The words "ke," "dula," and "le" are colored red, indicating they are pushing the prediction towards a negative sentiment, but their impact is outweighed by the stronger positive influence of other words.

Figure 12. The SHAP visualizations of a tweet with a positive sentiment.

In conclusion, our XAI methods effectively interpret the sentiments of isiZulu and Xhosa tweets, outperforming those in Sepedi, Sesotho, and Setswana. This could be influenced by the linguistic similarities between isiZulu and isiXhosa, as well as model performance in these languages. While our results show a general trend that higher model performance may lead to more accurate explanations, further analysis is required to confirm the existence of a direct correlation between model performance and explanation accuracy. Moreover, challenges arise when explaining less certain predictions, especially in languages where the model struggles. In particular, our approach identifies the keywords that influence sentiment, providing valuable insights into model decisions. Additionally, the effectiveness of our XAI methods extends to code-switched tweets, showcasing their potential in diverse linguistic environments. Lastly, XAI methods utilizing PLMs can be applied to any low-resourced African language that has been fine-tuned for sentiment analysis.

5.6. Human Assessment

One of the main goals in the XAI research field is to assist end-users in becoming more effective in the use of AI decision-support systems. The effectiveness of the human–AI interface can be measured by the AI's ability to support the human in achieving the task. This is also the case in the context of NLP systems. To evaluate the methods used for XAI in our sentiment analysis for low-resourced languages, we used an interactive human-centered feedback strategy to further assess the XAI methods. Each participant provided informed consent before taking part in the research. The results and explanations of two sets of participants are presented in the subsections below. The participants included experienced individuals with technical expertise who often use AI models (56.7%) and users who are familiar with AI models but do not use them (38.3%). A small group of participants (5%) opted not to participate in this study.

The feedback strategy used online questionnaires, with each question rated on a scale of 1 to 5 or as a binary choice of [Yes, No]. Participants were allowed to use the online sentiment analysis system, which generates sentiment predictions and explanations. They visualized their examples from the sentiment analysis system before completing the online questionnaires to rate the level of interpretation and explanation of the decision. The participants were asked to evaluate trust, transparency, reliability, and interpretability. The questions used in the study were built using Google Forms (The survey questions can be found here: https://forms.gle/oHHUw5ECwqjK1E66A (accessed on 12 November 2024). The results of 57 completed questionnaires are presented in Figures 13–16. These results

focused on answering questions based on the explanations generated by the LIME and SHAP methods.

5.6.1. Trustworthiness

Trust is a crucial factor in the context of XAI, especially in sentiment analysis applications, as users often want to understand and trust the decisions made by AI models. Four key questions were considered when asking participants to assess trust in XAI methods for sentiment analysis. The following questions were considered:

1. How confident do you feel in the accuracy of the sentiment analysis results provided by the XAI system?
2. How valuable did you find the explanations in helping you to trust the interpretations of the sentiment analysis results?
3. Did the XAI system's explanations positively influence any changes in your decision-making based on the sentiment analysis results?
4. How likely are you to use the XAI system for future sentiment analysis tasks, considering the explanations it provides?

The results that evaluated the trust factor are presented in Figure 13. The scale results indicate the distribution of responses regarding confidence in the accuracy of explanations provided by the XAI system. Each numerical value from 1 to 5 represents different levels of confidence (1—not confident; 2—slightly confident; 3—somewhat confident; 4—moderately confident; and 5—highly confident) or likelihood (1—not likely; 2—slightly likely; 3—somewhat likely; 4—moderately likely; and 5—highly likely).

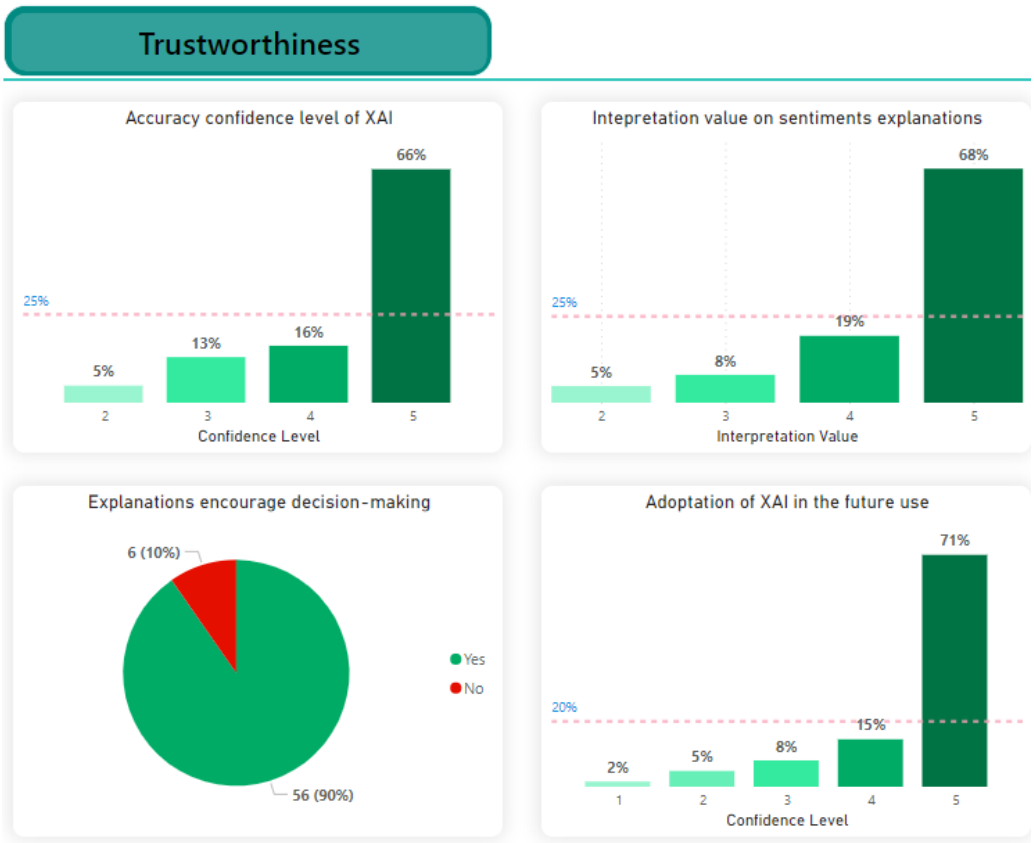


Figure 13. Results of the trustworthy factor of the XAI methods.

None of the participants indicated they had no confidence in the accuracy of the explanations provided by the XAI system. This suggests that no one dismissed the accuracy outright, which can be viewed as a positive sign. A substantial majority of participants (66%) displayed high confidence in the accuracy of explanations provided by the XAI

system. This reflects a generally positive outlook and a high level of trust in the system's performance. Although a significant portion expressed high confidence, there were still some participants (18%) who showed moderate confidence levels. Many of the participants had an understanding of AI models but did not use them.

The results of the scale indicate that a significant proportion of participants (68%) feel very confident in the value of explanations to trust and interpret the results of the sentiment analysis. A gradual increase in confidence levels is evident across the scale, with only 5% expressing slight confidence, indicating a strong overall positive sentiment towards the explanatory value of the XAI method. The distribution suggests that a majority of participants rely on and value these explanations for establishing trust in the sentiment analysis outcomes.

Furthermore, the results indicate a high level of influence from the XAI system's explanations on decision-making, with 90% of the participants reporting that the system's explanations did impact their decisions. This strong positive response suggests that the ability of the XAI system to provide transparent and interpretable explanations played a significant role in influencing participants' decision-making processes, highlighting the importance of effective XAI in enhancing trust and usability in sentiment analysis applications.

The scale results indicate a clear inclination towards using the XAI system for future sentiment analysis tasks, as 71% of participants expressed a high likelihood. A substantial majority of the participants, totaling 94%, expressed the likelihood of using the XAI methods for explanations (rating 4 or 5), suggesting strong potential adoption. However, there remains a smaller percentage of 6% (ratings 1 to 3) who could require further investigation regarding the explanations of the XAI system for general acceptance. Despite a minority expressing lower inclinations (2% in 'not likely'), the overall trend shows favorable acceptance and significant potential for adoption due to the perceived value of the explanations provided by the XAI system.

5.6.2. Transparency

Transparency promotes trust in the model. Trust is an emotional and cognitive component that determines how a system is perceived, either positively or negatively. Thus, when the decision-making process in the XAI model is thoroughly understood, the model becomes transparent. For this case, we evaluated the transparency based on two key questions:

1. Did the explanations provided by the XAI system help you understand why certain sentiments were identified in the text?
2. Did you encounter any challenges or difficulties while trying to understand the XAI system's generated explanations?

The results of the participants for the above-mentioned questions are presented in Figure 14.

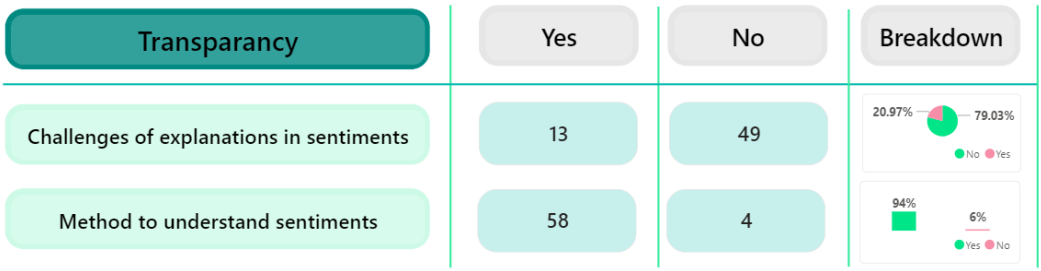


Figure 14. Results of the transparency factor of the XAI methods.

A significant majority, representing 79.03% of the participants, reported not facing any challenges or difficulties in interpreting the explanations of the XAI system for sentiment. This overwhelming result suggests that the explanations of the system may have been clear and easily understood by a large percentage of users. Nonetheless, the 20.97% of users who

experienced issues could suggest that the system's interpretability needs to be improved to accommodate a smaller but significant portion of users who found the explanations difficult to understand. Further enhancements to the XAI's explanation methods could be beneficial to address the concerns raised by this minority.

Lastly, the substantial 94% positive response to the question concerning whether the explanations provided by the XAI system helped to understand why certain sentiments were identified in the text indicates a high level of effectiveness and clarity in the system's explanations. This significant majority suggests that the interpretability features of XAI were successful in clarifying the reasoning behind the sentiment predictions of the vast majority of participants, highlighting a strong positive impact on comprehension and trust in the AI decision-making process. However, the relatively low 6% unfavorable reaction calls for further examination to identify specific areas of the visualization that need improvement or to address the requirements of users who may have experienced difficulties or found the explanations unsatisfactory, ultimately improving overall transparency and customer satisfaction.

5.6.3. Reliability

This section focused on evaluating the effectiveness and reliability of the XAI methods. For this, we evaluated the following two key questions:

1. How confident do you feel in the accuracy of the sentiment analysis results provided by the XAI system?
2. Were there any specific instances where you disagreed with the sentiment assigned by the XAI system, despite its explanations?

As indicated in Figure 15, the results for the reliability of the model provided by the XAI explanation methods show a range of confidence levels among participants.

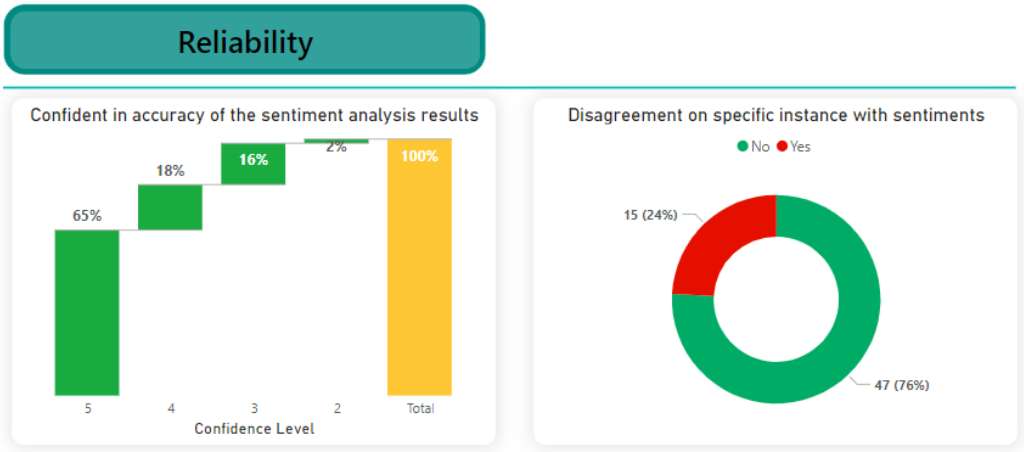


Figure 15. Results of the reliability factor of the XAI methods (based on LIME and SHAP visualization).

Most of the participants (65%) expressed high confidence in the precision of the sentiment analysis. However, a significant portion, 18%, also exhibited a fairly high level of confidence. A lower percentage, 16%, indicated moderate confidence, while only 2% showed slight confidence. None of the participants expressed no confidence in the accuracy of the XAI system sentiment analysis results.

The answers to the second question show that 76% of the respondents agreed with the sentiment that the XAI system assigned and the justifications that followed. However, 24% disagreed with the sentiment despite the explanations provided. This suggests a substantial level of trust, yet a notable portion of participants still encountered instances where they disagreed with the sentiment assigned by the XAI system, despite the explanations offered. Further explorations into these disagreements may provide valuable insights concerning

improving the reliability and accuracy of the XAI system for sentiment analysis in a low-resourced language context.

5.6.4. Interpretability

Interpretability is defined as the ability to explain to a human being the decision made by an AI model [5]. In this case, we further evaluated interpretability in the XAI methods by asking the participants three questions:

1. Did the explanations provided by the XAI system help you understand why certain sentiments were identified in the text?
2. Were there any instances where the XAI system's explanation of sentiment did not align with your interpretation?
3. Are there any challenges or difficulties you encountered while trying to interpret the XAI system's explanations of sentiment?

Figure 16 shows the results obtained.

Interpretability	Yes	No	Breakdown
XAI help with reason behind a sentiments	57	5	92% Yes No
Explanation alignment with sentiments interpretation	16	46	74% 26% No Yes
Challenges while interpreting sentiments	13	49	79% 21% No Yes

Figure 16. Results of the interpretability factor of the LIME and TransSHAP methods.

The results indicate a high level of agreement, with 92% of the participants answering 'Yes' and only 8% responding 'No' to the question about the effectiveness of the explanations provided by the XAI system in understanding the identified sentiments in the text. This substantial majority of positive responses suggests that the explanations provided by the XAI system significantly enhanced the participants' understanding of why specific sentiments were identified within the tweets. These findings strongly support the idea that the explanations generated by the XAI system played a crucial role in helping users understand the results of the sentiment analysis.

Additionally, the results indicate that 74% of participants reported instances where the XAI system's explanation of sentiment did not match their interpretation, while 26% stated otherwise. These results suggest that a significant majority found discrepancies between their understanding of sentiments and the explanations provided by the XAI system. Such disparities may indicate potential limitations in the system's ability to align with human interpretation, raising concerns about its reliability in accurately explaining sentiments. More research is required on the nature of these discrepancies to improve the reliability of the XAI system and bridge the gap between human and AI sentiment interpretations.

Finally, the findings show that a majority (79%) of participants did not encounter challenges when interpreting the explanations provided by the XAI system regarding sentiment analysis. However, 21% of participants acknowledged facing difficulties or obstacles when trying to understand the explanations of the XAI system. This distribution highlights a predominant ease in understanding the system's explanations for sentiment analysis, although a notable minority encountered challenges with certain aspects, indicating a need for potential improvements in explanation clarity or user guidance to enhance comprehension.

6. Limitations of the Study

We acknowledge a few limitations of this study. First, the availability of large, high-quality datasets for sentiment analysis in under-resourced African languages remains a

challenge. While the SAfriSenti corpus provided a valuable foundation, expanding the dataset to include more languages and domain-specific data would improve model robustness. Secondly, our approach did not fully address issues such as code-switching, which is common in African social media texts. Handling this phenomenon requires more advanced and specialized techniques for multilingual text processing. Thirdly, explainability methods such as LIME and SHAP may have limitations in fully capturing the complexities of transformer models, especially when dealing with complex linguistic structures or subtle sentiment expressions. Fourthly, the high computational cost of fine-tuning large PLMs like Afro-XLMR and SERENGETI can limit their accessibility in resource-constrained settings. Finally, limitations in the handling of sarcastic phrases and idiomatic expressions may lead to incorrect sentiment results in tweets processed by MPLM models.

7. Conclusions

This work presents a novel approach to sentiment analysis for resource-constrained African languages by integrating Afrocentric PLMs with XAI techniques. This integration not only improves sentiment prediction but also enhances the explainability of model decisions, providing more reliable and transparent AI systems. Although sentiment analysis tools and techniques are available in English, it is essential to extend coverage to other low-resourced African languages as well, ensuring that the models make informed decisions. We used the SAfriSenti corpus to perform model fine-tuning and sentiment classification on five low-resourced African languages. We modified LIME and TransSHAP to be suitable for transformer-based Afrocentric models. Although the models performed relatively well across the languages, in this sentiment analysis task the Afro-XLMR model outperformed them all, showing its adaptability to low-resourced languages. We used XAI tools to visualize the attention in the texts. We then used LIME and TransSHAP to visualize the tweets and their interpretations. While the Afro-XMLR model effectively interprets words conveying the sentiment, it provides inaccurate explanations in languages with lower sentiment performance. Furthermore, a feedback survey included a quantitative comparison of the explanations provided by TransSHAP and LIME, in addition to evaluating trustworthiness, reliability, transparency, and interoperability. The results of the survey pointed out that a significant majority of participants agreed with the decisions made by the XAI methods combined with the Afrocentric PLM model. Based on the outcomes generated by these XAI methods, we can conclude that they not only enhance the accuracy and interpretability of sentiment predictions but also promote understandability. This, in turn, cultivates trust and credibility in the decision-making process facilitated by Afro-XLMR as a sentiment analysis classifier. The use of XAI methods for sentiment explanations can be easily extended to other African languages.

In future work, we plan to extend our methods to handle challenges such as code-switching, dialect variations, and more complex explanation techniques, such as counterfactual visualizations to further improve model interpretability. Future work will explore handling code-switching in African languages and enhancing XAI techniques to address multilingual input. We also plan to address problems concerning the perturbation-based explanation process when dealing with textual data. Furthermore, we intend to investigate counter-factual visualizations, in which we alter various elements to assess the impact of the explanations. By extending the features of explanations from single words to longer textual units composed of words that are grammatically and semantically related, the explanations could be further improved. Finally, to make our approach more accessible and scalable in low-resourced settings, we will prioritize reducing computational costs through techniques such as model pruning, quantization, and knowledge distillation.

Author Contributions: Conceptualization, K.R.M. and M.P.; methodology, K.R.M.; software, K.R.M.; validation, K.R.M., M.P. and T.C.; formal analysis, K.R.M.; investigation, K.R.M.; resources, K.R.M.; data curation, K.R.M., M.P. and T.C.; writing—original draft preparation, K.R.M.; writing—review

and editing, M.P. and T.C.; visualization, K.R.M.; supervision, M.P. and T.C.; project administration, K.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with declarations and requirements of Twitter/X. The study was approved by the Human Research Ethics Committee (Non-Medical) at the University of the Witwatersrand (protocol number: HRECNMW21/10/07, issued on 14 October 2021), ensuring compliance with ethical standards in research.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. Additionally, we fully comply with Twitter's Developer Agreement and Policy, which governs the use, storage, and redistribution of tweet content.

Data Availability Statement: This data will not be commercialized. In line with Twitter's guidelines, only tweet IDs will be shared to allow other researchers to independently retrieve (rehydrate) data, ensuring compliance with ethical data sharing practices. Our Twitter dataset will be made publicly available and accessible by all NLP communities at https://github.com/NLPforLRLsProjects/SAfriSenti-Corpus (accessed on 12 November 2024).

Acknowledgments: The authors express their sincere gratitude to all the anonymous participants who helped to significantly improve the quality of work presented in this paper. Our work was supported by the National Research Foundation, Black Academics Advancement Programme, South Africa.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
LLM	Large Language Model
NLP	Natural Language Processing
PLM	Pre-trained Language Model
Afro-XLMR	Afro-centric XLM-RoBERTa Model
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	Shapley Additive Explanations
SAfriSenti	South African Sentiment Corpus
CNN	Convolutional Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
SOTA	State-of-the-Art
SMOTE	Synthetic Minority Over-sampling Technique

References

1. Ricardo, V.; Hossein, A.; Iolanda, L.; Madeline, B.; Virginia, D.; Sami, D.; Anna, F.; Daniela, L.S.; Max, T.; Francesco, F.N. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nat. Commun.* **2020**, *11*, 233.
2. Sharma, H.D.; Goyal, P. An Analysis of Sentiment: Methods, Applications, and Challenges. *Eng. Proc.* **2023**, *59*, 68. [CrossRef]
3. Enja, K.; Blaž, Š.; Nada, L.; Senja, P.; Marko, R.-Š. BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. In Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Association for Computational Linguistics, Online, 19 April 2021; pp. 16–21.
4. Fantozzi, P.; Naldi, M. The Explainability of Transformers: Current Status and Directions. *Computers* **2024**, *13*, 92. [CrossRef]
5. Arrieta, A.B.; Díaz-Rodríguez, N.; Ser, J.D.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
6. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
7. Amin, O.; Brown, B.; Stephen, B.; McArthur, S. A case-study led investigation of explainable AI (XAI) to support deployment of prognostics in industry. In Proceedings of the European Conference Of The PHM Society 2022, Prague, Czech Republic, 3 July–5 July 2022; Do, P., Michau, G., Ezhilarasu, C., Eds.; PHM Society: Lorne, Australia, 2022; pp. 9–20.

8. United Nations. Sustainable Development Goals: 17 Goals to Transform our World. 2022. Available online: https://www.un.org/sustainabledevelopment/sustainable-development-goals/ (accessed on 14 November 2024).
9. Loh, H.W.; Ooi, C.P.; Seoni, S.; Barua, P.D.; Molinari, F.; Acharya, U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput. Methods Programs Biomed.* **2022**, *226*, 107161. [CrossRef]
10. Kumar, P.; Hota, L.; Tikkiwal, V.A.; Kumar, A. Analysing Forecasting of Stock Prices: An Explainable AI Approach. *Procedia Comput. Sci.* **2024**, *235*, 2009–2016. [CrossRef]
11. Schoonderwoerd, T.A.; Jorritsma, W.; Neerincx, M.A.; Van Den Bosch, K. Human-centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *Int. J. Hum.-Comput. Stud.* **2021**, *154*, 1–25. [CrossRef]
12. Song, M. A Study on Explainable Artificial Intelligence-based Sentimental Analysis System Model. *Int. J. Internet Broadcast. Commun.* **2022**, *14*, 142–151.
13. Mayur, W.; Annavarapu, R.; Chaitanya, K. A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780.
14. Ronny, M.K.; Turgay, C.; Mpho, R. Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access* **2023**, *11*, 15996–16020.
15. Anh, L.T.; David, M.; Yasuhide, M.; Tomoko, O. Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. In Proceedings of the 12th Workshop on Asian Language Resources (ALR12), Osaka, Japan, 12 December 2016; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 123–131.
16. Ife, A.; AbdelRahim, E.; Muhammad, A.; Alcides, A.I. SERENGETI: Massively Multilingual Language Models for Africa. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 1498–1537.
17. Alabi, J.O.; Ifeoluwa, A.D.; Marius, M.; Dietrich, K. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 4336–4349.
18. Dossou, B.F.P.; Tonja, A.L.; Yousuf, O.; Osei, S.; Oppong, A.; Shode, I.; Awoyomi, O.O.; Emezue, C. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), Abu Dhabi, United Arab Emirates, 7 December 2022; pp. 52–64.
19. Ronny, M.K.; Mpho, R.; Turguy, C. Investigating Sentiment-Bearing Words- and Emoji-based Distant Supervision Approaches for Sentiment Analysis. In Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023), Dubrovnik, Croatia, 2–6 May 2023; pp. 115–125.
20. Aguero-Torales, M.M.; Salas, J.I.A.; Lopez-Herrera, A.G. Deep learning and multilingual sentiment analysis on social media data: An overview. *Appl. Soft Comput.* **2021**, *107*, 107–373. [CrossRef]
21. Jacob, D.; Ming-Wei, C.; Kenton, L.; Kristina, T. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (Long and Short Papers); pp. 4171–4186.
22. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
23. Alexis, C.; Kartikay, K.; Naman, G.; Vishrav, C.; Guillaume, W.; Francisco, G.; Edouard, G.; Myle, O.; Luke, Z.; Veselin, S. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8440–8451.
24. Kelechi, O.; Yuxin, Z.; Jimmy, L. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In Proceedings of the 1st Workshop on Multilingual Representation Learning, Punta Cana, Dominican Republic, 11 November 2021; pp. 116–126.
25. Bacco, L.; Cimino, A.; Dell'Orletta, F.; Merone, M. Explainable Sentiment Analysis: A Hierarchical Transformer-Based Extractive Summarization Approach. *Electronics* **2021**, *10*, 2195. [CrossRef]
26. Mahendhiran, P.; Subramanian, K. Deep Learning Techniques for Polarity Classification in Multimodal Sentiment Analysis. *Int. J. Inf. Technol. Decis. Mak.* **2018**, *17*, 883–910. [CrossRef]
27. Aliyu, Y.; Sarlan, A.; Danyaro, K.U.; Rahman, A.S.B.A.; Abdullahi, M. Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. *IEEE Access* **2024**, *12*, 66883–66909. [CrossRef]
28. Mahendhiran, P.D.; Subramanian, K. CLSA-CapsNet: Dependency based concept level sentiment analysis for text. *J. Intell. Fuzzy Syst.* **2022**, *43*, 107–123. [CrossRef]
29. Arunkumar, P.; Chandramathi, S.; Kannimuthu, S. Sentiment analysis-based framework for assessing internet telemedicine videos. *Int. J. Data Anal. Tech. Strateg.* **2019**, *11*, 328–336. [CrossRef]
30. Ronny, M.K.; Tim, S. AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa. In *Artificial Intelligence Research*; Springer: Cham, Switzerland, 2022; pp. 309–322.
31. Lecturer, M.O.; Lecturer, K.S.; Kaliyaperumal, D.K. Sentiment analysis for afaan oromoo using combined convolutional neural network and bidirectional long short-term memory. *Int. J. Adv. Res. Eng. Technol.* **2020**, *11*, 101–112.
32. Suri, V.; Arora, B. A Review on Sentiment Analysis in Different Language. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1–9.

33. Kelly, S.; Kaye, S.; Oviedo-Trespalacios, O. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telemat. Inform.* **2023**, *77*, 101925. [CrossRef]
34. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl.-Based Syst.* **2023**, *263*, 110273. [CrossRef]
35. Kun, Q.; Marina, D.; Yannis, K.; Ban, K.; Erick, O.; Lucian, P.; Yunyao, L. XNLP: A Living Survey for XAI Research in Natural Language Processing. In Proceedings of the 26th International Conference on Intelligent User Interfaces—Companion, College Station, TX, USA, 14–17 April 2021; IUI '21 Companion; pp. 78–80.
36. Liu, S.; Franck, L.; Supriyo, C.; Tarek, A. On Exploring Attention-based Explanation for Transformer Models in Text Classification. In Proceedings of the IEEE International Conference on Big Data, Orlando, FL, USA, 15–18 December 2021; pp. 1193–1203.
37. Park, S.; Lee, J. LIME: Weakly-Supervised Text Classification without Seeds. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 1083–1088.
38. Bodria, F.; Panisson, A.; Perotti, A.; Piaggesi, S. Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis. In Proceedings of the Sistemi Evoluti per Basi di Dati, Sud Sardegna, Italy, 21–24 June 2020.
39. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
40. Lundberg, S.M.; Lee, Su. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
41. Arwa, D.; Kawther, S.; Kia, D.; Mandar, G.; Erik, C.; Amir, H. Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis. *IEEE Trans. Affect. Comput.* **2023**, *15*, 837–846.
42. Library of Congress. ISO 639-2 Language Code List. Available online: https://www.loc.gov/standards/iso639-2/php/code_list.php (accessed on 14 November 2024).
43. Ronny, M.; Tim, S. A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, Marseille, France, 24–25 June 2022; pp. 70–77.
44. Muhammad, S.; Abdulmumin, I.; Ayele, A.; Ousidhoum, N.; Adelani, D.; Yimam, S.; Ahmad, I.; Beloucif, M.; Mohammad, S.; Ruder, S.; et al. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 13968–13981. [CrossRef]
45. Arras, L.; Montavon, G.; Müller, K.R.; Samek, W. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Copenhagen, Denmark, 8 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 159–168.
46. Jesse, V. A Multiscale Visualization of Attention in the Transformer Model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, 28 July–2 August 2019; pp. 37–42.
47. Saleem, R.; Yuan, B.; Kurugollu, F.; Anjum, A.; Liu, L. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing* **2022**, *513*, 165–180. [CrossRef]
48. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the ICLR, Addis Ababa, Ethiopia, 30 April 2020.
49. Lee, J.; Tang, R.; Lin, J.J. What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. *arXiv* **2019**, arXiv:1911.03090.
50. Hadwan, M.; Al-Sarem, M.; Saeed, F.; Al-Hagery, M.A. An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique. *Appl. Sci.* **2022**, *12*, 5547. [CrossRef]