Article

Not peer-reviewed version

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

Ronny Mabokela * , Mpho Primus , Turgay Celik

Posted Date: 3 September 2024

doi: 10.20944/preprints202409.0253.v1

Keywords: Explainable AI; sentiment analysis; African languages; Africa-Centric models; pre-trained models; transformer models

Article

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

Ronny Mabokela *, Mpho Primus and Turgay Celik

Affiliation 1

* Correspondence: krmabokela@gmail.com

Abstract: Sentiment analysis is a pivotal tool for gauging the public's perception and understanding of human communication across digital social media platforms. However, due to linguistic complexities and limited resources, sentiment analysis is not well-represented in many African languages. While benchmark Africa-Centric Pre-trained Language Models (PLMs) have been developed for various Natural Language Processing (NLP) tasks, their applications in eXplainable Artificial Intelligence (XAI) remain unexplored. In this study, we introduce a novel approach that combines Africa-centric PLMs with XAI techniques for sentiment analysis. We demonstrate that applying attention mechanisms and visualisation techniques improves the transformer-based model's *transparency*, *trustworthiness*, and *decision-making* abilities when making sentiment predictions. We then employ the ***SAfriSenti***—a multilingual sentiment corpus for South African under-resourced languages. We use the corpus to perform various sentiment analysis experiments and also enable comprehensive evaluations, comparing the performance of Africa-centric models against mainstream PLMs. The Afro-XLMR model outperformed all models and achieved an average F1-score performance of 71.04% across the five tested languages and the lowest error rate among the evaluated models. Additionally, we incorporated techniques like Local Interpretive Model-Agnostic Interpretation (LIME) and Shapley Additive Interpretation (SHAP) in the sentiment classifier's output to enhance the Afro-XLMR model's *interpretability* and *explainability*. As a result, the use of XAI strategies ensures that sentiment predictions are not only *accurate* and *interpretable* but also *understandable*, fostering *trust* and *reliability* in the decision-making of AI-driven NLP technologies, particularly in the context of African languages.

Keywords: Explainable AI; sentiment analysis; African languages; Africa-Centric models; pre-trained models; transformer models

1. Introduction

Artificial intelligence (AI) has become increasingly popular recently, 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 emerging recently. This includes large language models (LLMs) based on transformer models, ensemble methods, and Deep Neural Networks (DNNs) [3,4]. Nevertheless, 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 transparency 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 exists in integrating human factors into AI-generated explanations, making them more understandable and performance-enhancing in NLP for African languages. XAI enables users to develop trust in NLP applications like sentiment analysis, creating a positive feedback loop that continually improves the NLP model's performance.

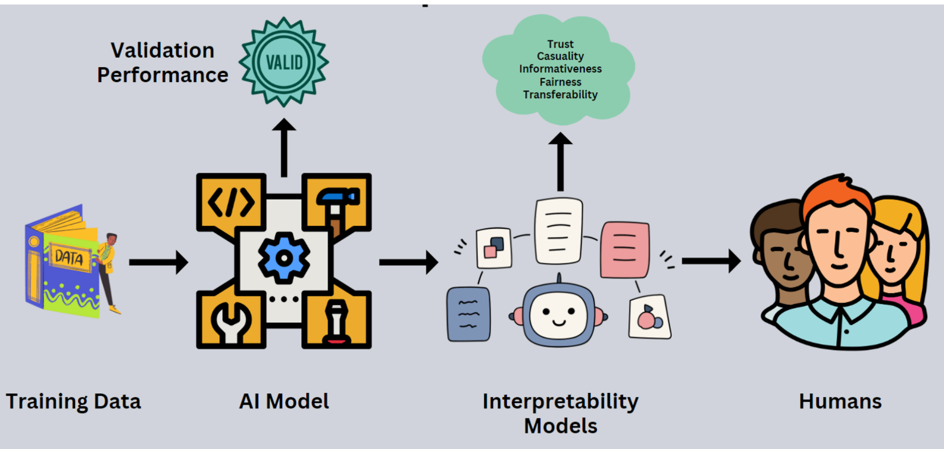


Figure 1. A step-by-step general purpose framework for XAI systems.

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]. Most recently, sentiment analysis research has extended its focus from high-resourced languages to low-resourced languages, such as 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 labelled data and resources, thus presenting unique challenges for existing sentiment analysis models. Moreover, due to the increased use of social media, particularly Twitter, 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, multilingual and cross-lingual approaches [14], utilizing machine learning and DNN approaches [20]. Lately, we have witnessed the development of sentiment datasets for low-resource languages, together with multilingual 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 Afro-centric 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 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, Africa-centric 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 utilise features to classify text into sentiment categories. This lack of transparency hinders our ability to trust and interpret their decisions.

Fortunately, XAI methods offer a solution [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 can apply XAI techniques like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive exPlanations), 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 Afro-centric PLMs adapted specifically for low-resource 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.

This study introduces an approach to sentiment analysis that combines Africa-centric PLMs with XAI techniques. We explore the applicability of LIME and SHAP, specifically focusing on their ability to generate explanations for transformer models like Afro-centric PLMs. To strengthen our multilingual sentiment analysis models for five under-resourced African languages: *Sepedi*, *Setswana*, *Sesotho*, *isiXhosa*, and *isiZulu*, we strategically employ the SAfriSenti corpus. This corpus is crucial for fine-tuning our models and evaluating their performance on sentiment analysis tasks. These languages are widely spoken in South Africa and extend to neighboring countries such as Botswana, Lesotho, Eswatini (formerly Swaziland), Mozambique, and Zimbabwe.

The main contributions of our study are summarised as follows:

- We propose a novel hybrid approach that integrates Africa-centric multilingual PLMs with XAI techniques. This integration allows us to apply the sentiment analysis capabilities of Afro-centric PLMs while simultaneously incorporating XAI methods to explain their decision-making processes for improved transparency and trust.
- Our approach utilises fine-tuned benchmark Africa-centric PLMs specifically designed for African languages. This choice capitalises on their understanding of linguistic nuances in these languages, potentially leading to superior sentiment analysis performance compared to mainstream PLMs.
- By incorporating attention mechanisms and visualization techniques, we enhance the transparency of the Africa-centric sentiment analysis model. This allows users to understand which parts of the input text the model focuses on when making sentiment predictions, fostering trust in its decision-making process.
- We demonstrate that incorporating LIME and SHAP techniques into the sentiment classifier's output enhances the model's interpretability and explainability.
- We also show that by leveraging XAI strategies, the study ensures that the model's sentiment predictions are accurately interpretable and understandable. Furthermore, the feedback survey shows that many of the participants are in agreement with the results of the models and XAI explanations.

In the next section, we present the related work and the latest approaches of other researchers for XAI methods for sentiment analysis concerning explainability in transformer-based PLMs. Section 3 presents our sentiment datasets. In Section 4, we provide our research on XAI techniques for sentiment analysis through the adaptation of Afrocentric PLMs. Section 5 describes the experimental setup for our study. The results and discussion are described in Section 5. We conclude our work and suggest further steps in Section 6.

2. Related Studies

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

2.1. Africa-Centric Pre-Trained Language Models

Since the introduction of the BERT model [21], XLM-R, and RoBERTa [23], PLM 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 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-resource 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]. [24] developed the AfriBERTa model for 11 African languages. [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 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 like the SERENGETI model [16] hold promise for broad applicability across African languages due to its massive 517-language coverage, its effectiveness and the explainability of its predictions using XAI techniques remain unexplored.

2.2. Explainable AI for Sentiment Analysis

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 [26]. The human interpretation aspect of the explanation holds significant value. If the user of the application fails to understand it, its applicability to the end user loses its purpose. Therefore, XAI methods are considered to assist domain experts in creating technically sound and human-interpretable explanations. XAI methods aim to make models intrinsically interpretable and transparent for blackbox models [27]. 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 visualisation are employed to provide insights into which parts of the input text contribute most to the sentiment analysis [12,28]. [29] introduced an attention-based explanation method for sentiment analysis using BERT. They demonstrated how attention scores reveal which parts of the input text contributed to the model's sentiment prediction. [30] 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. [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. [31] 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 behaviour of the model.

[32] introduced LIME, a model-agnostic method for generating locally faithful explanations applied to sentiment analysis. Researchers have utilized LIME to perturb input text instances and learn surrogate models that highlight the influential words affecting sentiment predictions [4]. LIME has proven effective in explaining the decisions of sentiment analysis models such as BERT and LSTM. [33] 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 behaviour [12,28]. Combining LIME and SHAP is a hybrid approach that leverages the strengths of both methods for interpreting sentiment predictions. This study emphasises the value of cross-method validation to ensure robust explanations [7,34]. Hence, for this study, we explore a combination of these XAI methods to achieve the objectives of our study.

3. Datasets

Table 1 illustrates the statistical distribution of tweets across five South African languages. These datasets are categorised 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 platform was changed to its official name, X [1]. In this study, we still refer to the textual information of the platform as tweets.

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

Language	**Positive**		**Negative**		**Neutral**		**Total**
(ISO 639)	**Pos**	**%**	**Neg**	**%**	**Neu**	**%**	
Sepedi (nso)	5,153	48%	3,270	30%	2,355	22%	10,778
Setswana (tsn)	3,932	51%	2,150	28%	1,590	21%	7,672
Sesotho (sot)	3,050	48%	2,024	32%	1,241	20%	6,314
isiXhosa (xho)	6,657	25.79%	12,125	48.10%	6,421	25.47%	25,203
isiZulu (zul)	19,252	42.49%	22,400	49.44%	3,378	7.45%	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 data sets 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 Africa 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*, *Sesotho*, including *English*. The *SAfriSenti* corpus was manually annotated by experts and native speakers. Each tweet in the corpus has undergone the data cleaning and preprocessing steps where noise, URLs, meaningless tweets were removed [35]. We replaced all @mentions by @user for data protection purposes [36]. 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 API to collect this dataset, following the approach presented in [35]. This sentiment dataset was collected from Twitter and contains over 50,000 tweets in the isiZulu and isiXhosa languages. To automate sentiment labelling 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 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% total tweets. The dataset also includes code-switched English words.

[1] X is the new name of the social media company and platform formerly known as Twitter

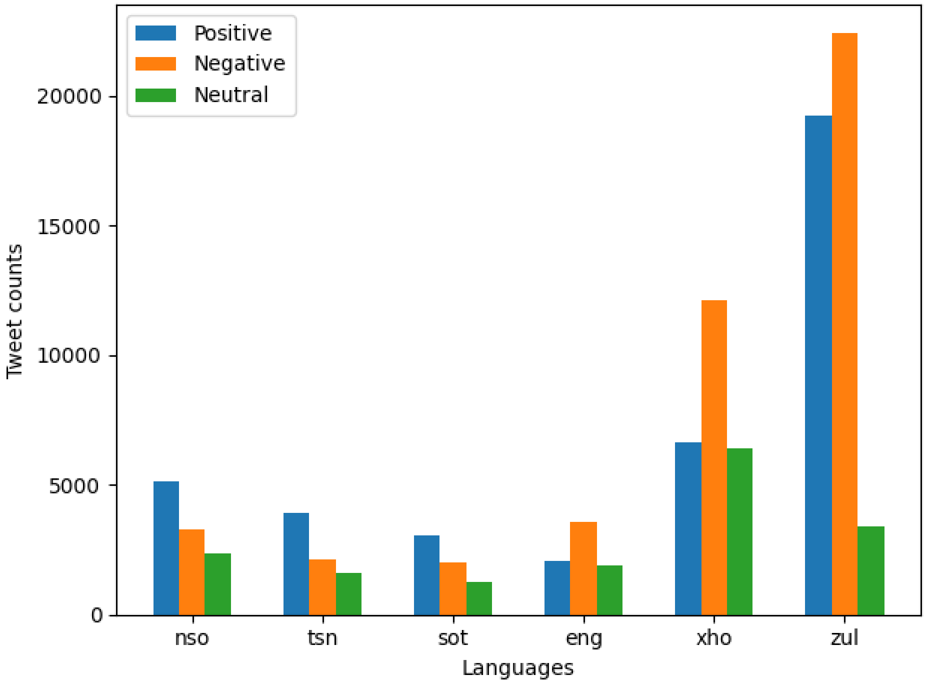


Figure 2. Distribution of Tweets Across Languages. We show the graphical comparison of the various tweets across the 5 Languages for sentiment analysis.

4. Proposed Methods for XAI for Sentiment Analysis

In this section, we describe our proposed XAI approaches for sentiment analysis using transformer-based models, as shown in Figure 3.

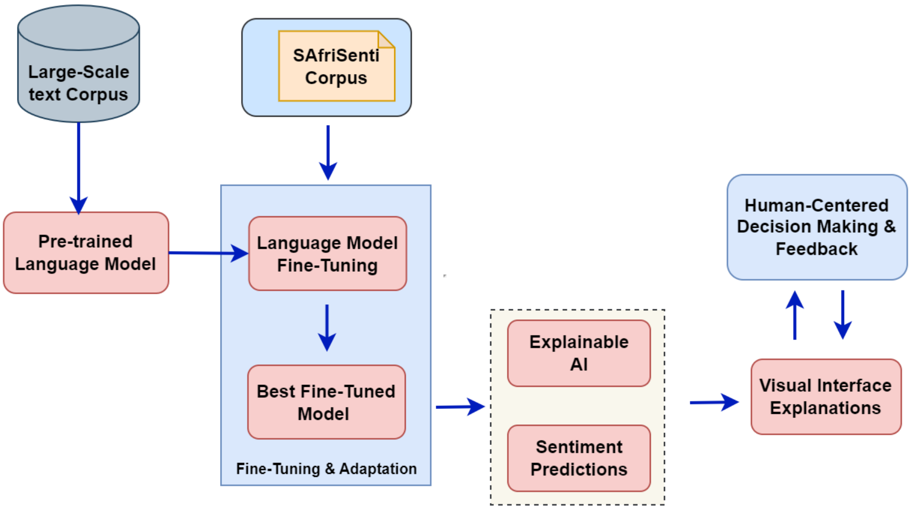

Large-Scale text Corpus
SAfriSenti Corpus
Pre-trained Language Model
Language Model Fine-Tuning
Best Fine-Tuned Model
Fine-Tuning & Adaptation
Explainable AI
Sentiment Predictions
Human-Centered Decision Making & Feedback
Visual Interface Explanations

Figure 3. Overview of Explainable AI Framework for Sentiment Analysis.

4.1. Overview of XAI for Sentiment Analysis

Figure 2 provides an overview of the system framework dedicated to leveraging PLMs for XAI methods in sentiment analysis. We employ the SAfriSenti corpus presented in Figure 2. 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 labelled portion of our sentiment dataset. Furthermore, 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. Model Descriptions

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:

- **mBERT:** Multilingual BERT is a multilingual version of BERT pre-trained in the top 104 languages with the largest Wikipedia using a masked language modelling (MLM) objective and the next-sentence prediction task [21]. We fine-tune the `bert-base-multilingual-cased` model with 172M model parameters by adding a linear classification layer on top of the pre-trained transformer model.
- **XLM-R (XLM-RoBERTa):** XLM-R 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.
- **AfroLM:** AfroLM is a multilingual language model that has been pre-trained from scratch on African languages using a novel self-active learning framework [18]. It is the only available SOTA Transformer model that has been pre-trained with 23 African languages, including Setswana, isiXhosa, and isiZulu in our case. AfroLM stands out for its efficiency. Trained on a considerably smaller dataset than existing models, it still surpasses many multilingual PLMs in various NLP tasks.
- **Afro-XLMR** Researchers created Afro-XLMR by adapting the XLM-R large model using MLM (Masked Language Modeling) on 17 African languages. This covers major language families like Sesotho, isiXhosa, and isiZulu, along with 3 high-resource languages including English [17]. Afro-XLMR is multilingual adaptive fine-tuning that allows multilingual adaptation and preserves downstream performance on both high- and low-resourced languages. 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 42GB of data comprising a multi-domain from religious, news, government documents, health documents, and existing Wikipedia corpora [16]. The pretraining data covers 517 African languages and the 10 most spoken languages globally. This model was pre-trained on both Electra style [37] as well as XLM-R style architecture. Electra utilises the multilingual replaced token detection (MRTD) objective during training. The model has 12 layers and 12 attention heads. SERENGETI model has significantly outperformed AfriBERTa, XLMR, mBERT, and Afro-XLMR on some NLP tasks.

4.3. 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 [38]. Hence, we utilize XAI methods for local explainability, which specializes 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 visualisation [27]. To calculate the attention per token, we need the weights for the encoder-decoder attention layers. However, attention serves a purpose beyond just optimisation; it also plays an essential role in expanding the context of the transformer-based language model [39]. We use **LIME**[2] and **TransSHAP**[3] (i.e., a version for transformer-based models) together with tools such as **Bertviz** to generate visualisations for further explanations [33]. The most widely used perturbation-based explanation methods are LIME with a wrapper function to support the Transformer-based model, and SHAP [5,34]. For the SHAP to work effectively with BERT-like models, particularly the Kernel SHAP, a function classifier that returns probabilities [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 [40]. The SHAP value of a particular feature for a data point is described by the following Equations (1) and (2):

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

where;

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

where x is the vector of feature values of tweet (which needs to be explained) and term S denotes a subset of input features and Shapley value 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 is used to make individual tweet classification predictions more understandable [32]. LIME achieves this by generating perturbed versions of the original tweet and evaluating their impact on the model's prediction [40]. The process of creating local surrogate models with interpretability constraints can be expressed as follows (Equation 3):

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

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 [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% favoured LIME, 42.1% for TransSHAP, and only 34.2% for 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%. 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. Additionally, to ensure that model output aligns with human values, we incorporated a feedback process involving human input through a survey.

[2] https://github.com/shap/shap
[3] https://github.com/enjakokalj/TransSHAP

5. Experimental Results and Explanations

In this section, we present the experiments and results of sentiment analysis systems. Furthermore, we explain the attention mechanisms and visualization outcomes using XAI techniques such as LIME and TransSHAP.

5.1. Experimental Setup

In this study, we conducted training and fine-tuning procedures on four PLMs for sentiment analysis on a set of five sentiment datasets. The data for each language are made up of the 80% and 20% data partition for training and testing, respectively. We perform model fine-tuning by considering the final hidden vectors of the first special token as the aggregate input sentence representation and then passing them onto the softmax classification layer to get the predictions. Our PLMs use hyperparameters like the AdamW optimizer with a 1×10^{-4} initial learning rate, 10 epochs, and a batch size of 16. All training experiments were performed using the HuggingFace Transformers library. We used paid Google Colab Pro services to run all of our experiments.

5.2. Performance Results

Table 2 demonstrates that the performance of the models varies significantly across different languages. The choice of language model has a significant impact on sentiment analysis results for different African languages. We also report on the F1-score of our models. The F1-score is a metric used to evaluate the performance of machine learning models, particularly in classification tasks [41]. It combines two essential aspects of model performance: precision and recall. 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 (TP): The number of tweets correctly labelled as having the positive sentiment.
- True Negatives (TN): The number of tweets correctly labelled as not having the positive sentiment.
- False Positives (FP): The number of items incorrectly labelled as positive sentiment.
- False Negatives (FN): The number of items incorrectly labelled as not having a positive sentiment.

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

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

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

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.

Table 2. 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). 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, further 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. Here, it is also observed that the Afro-centric PLMs also perform significantly better than the mBERT and XLM-R models in general.

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 shows its best performance in isiZulu, but still falls short of Afro-XLMR. XLM-R and SERENGETI also perform well. The results suggest that models specifically trained on African languages, such as Afro-XLMR, tend to perform better than more general models such as mBERT or XLM-R. This highlights the importance of developing language models tailored to specific linguistic contexts, especially for underrepresented languages. Our results are comparable to those of previous work [18,36]. The relatively strong performance of SERENGETI, AfroLM AND Afro-XMLR across different languages suggests that it might be a good choice for general-purpose language processing tasks in South Africa.

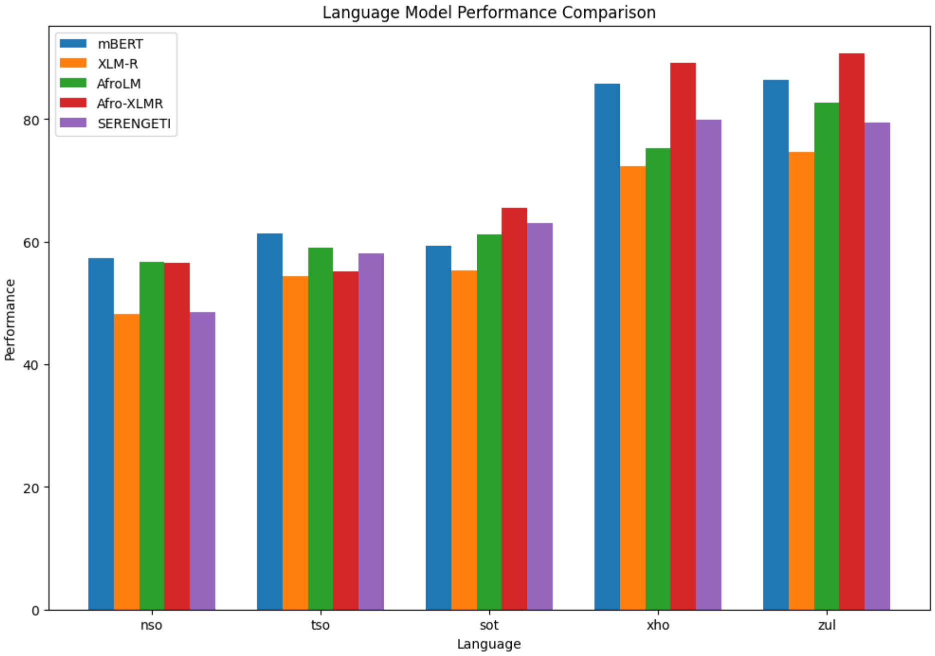


Figure 4. Overview of Explainable AI Framework for Sentiment Analysis.

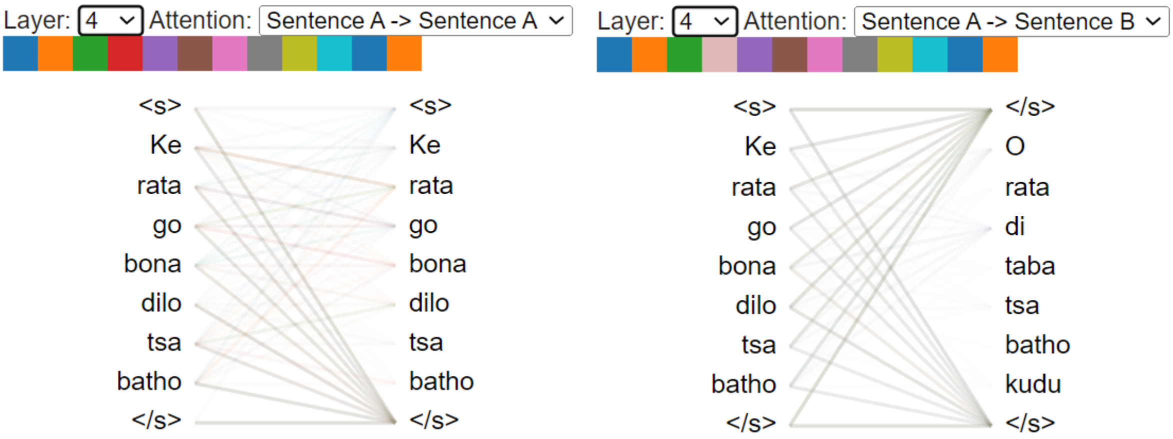
5.3. Attention Explanations

Attention allows the model to weigh the importance of different words in a sentence when understanding the meaning of a particular word. It is like 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.



(a) Self-attention.

(b) Attention between two sentences.

Figure 5. Comparison of attention visualizations.

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.

5.4. Sentiment Decision 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 [2,6]. We can use the LIME method to create visualizations such as the one in Figure 6, which helps us to understand which words in parts of the tweets have the greatest influence on the final prediction of the model. The model has analyzed the text given and determined that it has a 57% probability of being negative and a 43% probability of being positive. We can observe from the example ”Ke dula ke lapile and I am annoyed” /*I am always tired and I am annoyed*/ that the model is performing successfully and can obtain the feeling that the tweet is negative.

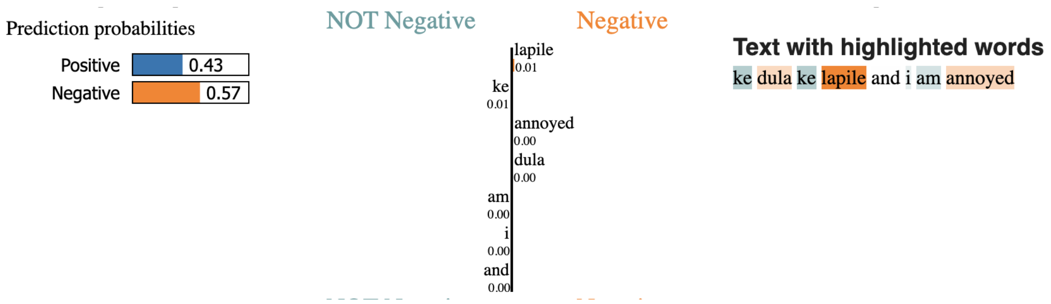


Figure 6. The LIME visualisations of a negative tweet in Sepedi mixed with English.

It 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 consideration of relevant words such as "lapile," "tired," and "annoyed" (orange), which indicate the negative sentiment of this tweet. Furthermore, the model also indicates 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. In this case, the model seems to be associating the words "lapile" and "ke" with negative sentiment, even though "ke" might not be inherently negative in all contexts.

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

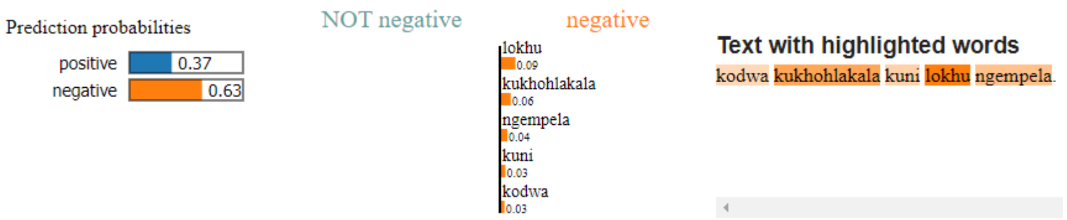


Figure 7. The LIME visualisations of a negative tweet in isiZulu language.

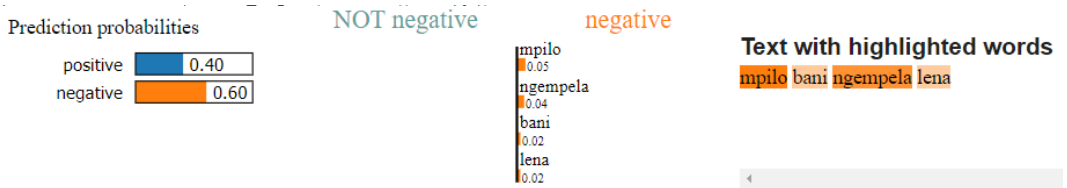


Figure 8. The LIME visualisations of a negative tweet in isiZulu language.

The model incorrectly classified the tweet in Figure 9 as "negative," despite the original tweet expressing a 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's 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.

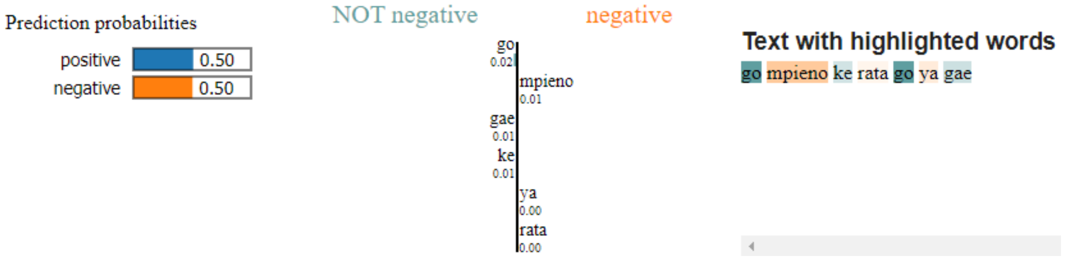
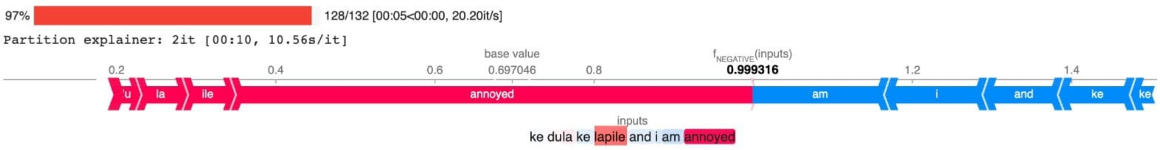


Figure 9. The LIME visualisations of an incorrect prediction in Setswana language.

The visualisation with the SHAP method for the same sentence in Figure 6 is illustrated in Figure 10. We can see that the features with the strongest impact on the prediction correspond to longer arrows that point in the direction of the predicted class, which is negative. This confirms that a model with higher accuracy tends to generate more reliable explanations. When the model accurately predicts the results, the explanations it provides for those predictions are more likely to be reliable and meaningful. This represents the average prediction of the model across the entire dataset. In this example, the base value leans slightly towards negative sentiment (0.697). This is the model's prediction for the specific input text "ke dula ke lapile and i am annoyed". The output value is almost 1.0, indicating a strong negative sentiment prediction. Each word in the input text is a feature. The plot shows the impact of each word on the model's prediction, moving it away from the base value towards the output value. The color and length of the horizontal bar for each feature (word) indicate the magnitude and direction of its impact. Red bars indicate features that push the prediction towards negative sentiment, while blue bars push toward positive. The longer the bars, the stronger the impact.



```
97%  128/132 [00:05<00:00, 20.20it/s]
Partition explainer: 2it [00:10, 10.56s/it]
```

Figure 10. The SHAP visualisations 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 (0.640081) compared to positive (0.36091). The words "lena" and "mpilo" contribute slightly towards the positive sentiment, while "ngempela" strongly pushes the prediction towards the 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 indicates the average prediction in the absence of any input features.

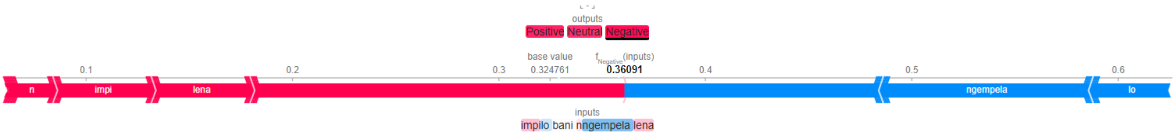


Figure 11. The SHAP visualisations of a tweet with negative sentiment.

In Figure 12, SHAP results of the text in Sepedi "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 0.3494 is 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, like "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.

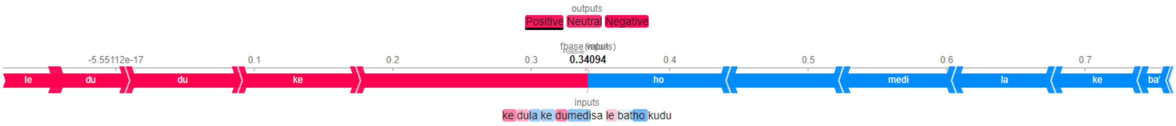


Figure 12. The SHAP visualisations of a tweet with a positive sentiment.

In conclusion, our XAI methods effectively interpret the sentiments of the Xhosa and isiZulu tweets, outperforming the Sepedi, Sesotho, and Setswana tweets, suggesting a correlation between model performance and explanation accuracy. However, 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 utilising PLMs can be applied to any African language that has been fine-tuned for sentiment analysis.

5.5. 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. In the context of NLP systems, this is also the case. 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 evaluate the XAI methods. The results and explanations of two sets of participants are presented in the below sub-sections. The participants, including experienced participants with technical expertise who often use the AI models (56.7%) and users who are familiar with AI models but do not use them (38.3%), were involved in this study. Only three (3) participants (5%) opted not to participate in this present study.

The feedback strategy uses online questionnaires, each question being rated on a scale of 1 to 5 or as a binary choice of [Yes, No]. Participants are allowed to use the online sentiment analysis system, which generates sentiment predictions and explanations. They do so by visualising 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 are asked to evaluate trust, transparency, reliability, and interpretability. The questions used in this study were built using GoogleForm [4]. 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.

[4] The survey questions are found here: https://forms.gle/oHHUw5ECwqjK1E66A

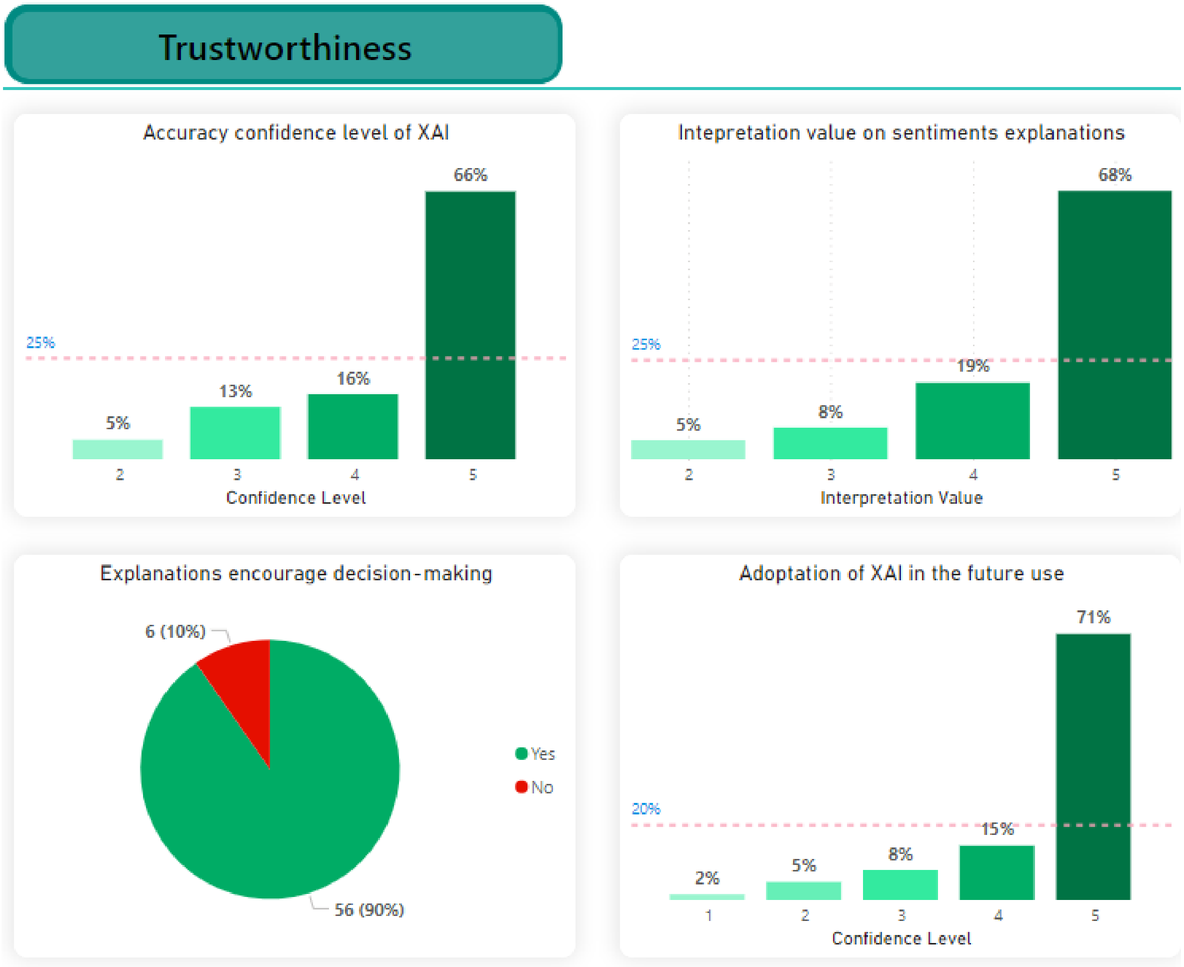
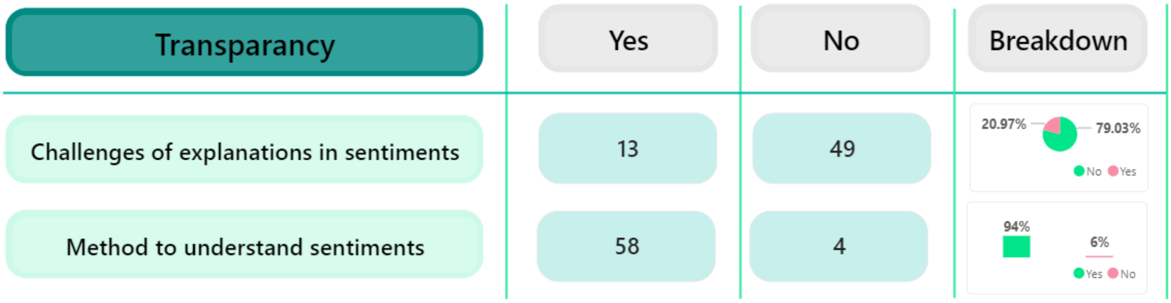


Figure 13. The results of the trustworthy factor of the XAI methods.



Transparancy	Yes	No	Breakdown
Challenges of explanations in sentiments	13	49	20.97% 79.03% ● No ● Yes
Method to understand sentiments	58	4	94% 6% ● Yes ● No

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

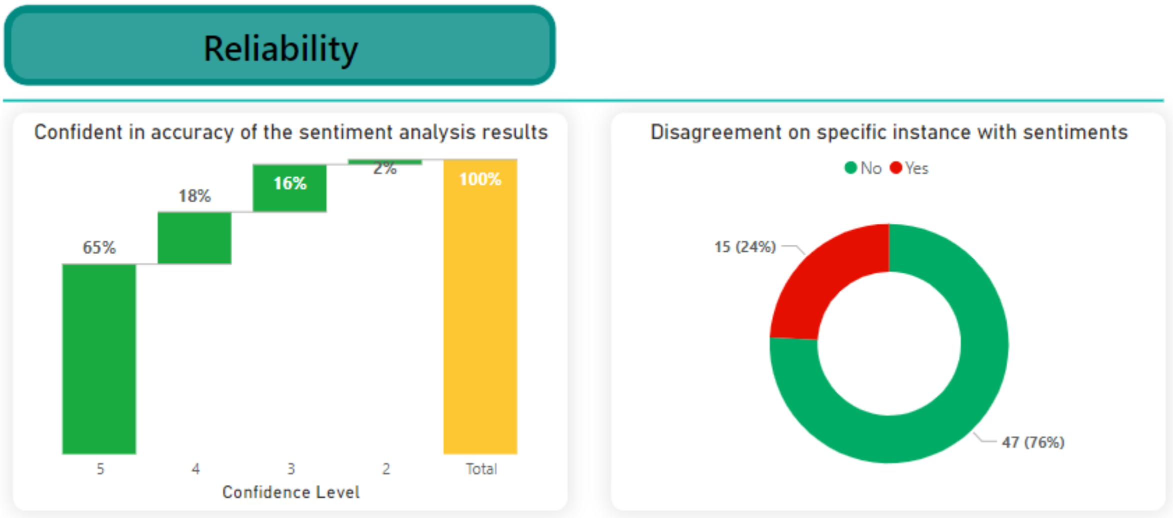
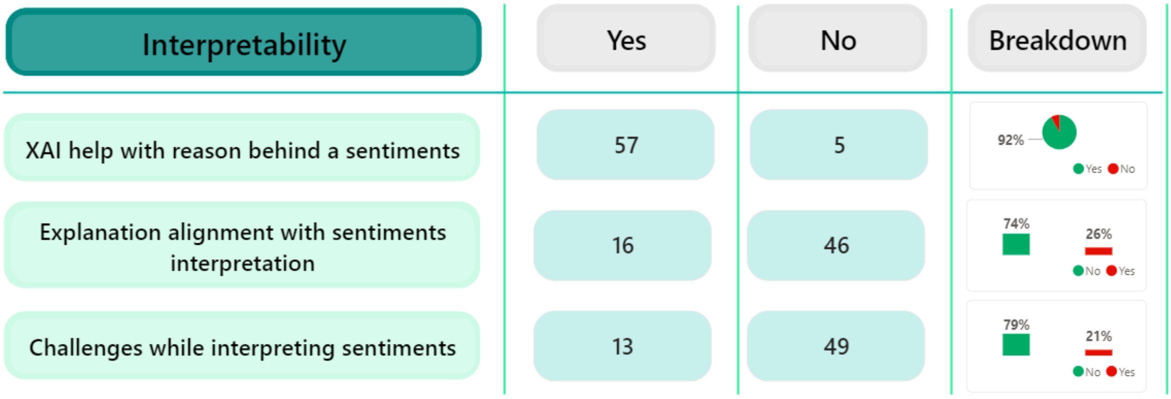


Figure 15. The results of the reliability factor of the XAI methods. These were based on LIME and SHAP visualisations.



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. The results of the interpretability factor of the LIME and TransSHAP methods. The results presented evaluated XAI methods such as LIME and TransSHAP.

5.5.1. 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 do you find the explanations in helping you trust and interpret 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 - Moderate confident and 5 - Highly confident) or (1 - Not likely, 2 - Slightly likely, 3 - Somewhat likely, 4 - Moderate likely and 5 - Highly likely).

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 outright dismissed the accuracy, 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 a 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 favourable acceptance and significant potential for adoption due to the perceived value of the explanations provided by the XAI system.

5.5.2. 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. *Were there any challenges or difficulties you encountered while trying to interpret the XAI system's explanations of sentiment?*
2. *Did the explanations provided by the XAI system help you understand why certain sentiments were identified in the text?*

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

The significant majority, which represented 79.03% of the participants reported not facing any challenges or difficulties in interpreting the explanations of the XAI system of 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 huge 94% positive response that states that 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% unfavourable reaction calls for more examination to identify particular areas of the visualisation for improvement or to meet the requirements of users who may have had difficulties or considered the explanations unsatisfactory, to improve general transparency and customer satisfaction.

5.5.3. Reliability

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

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.

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. In particular, a segment, representing 0%, expressed no confidence in the accuracy of the XAI system sentiment analysis results. Furthermore, the answers to the other 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 held a differing sentiment judgment from the XAI system, despite the explanations offered. Further exploration into these disagreements may provide valuable insights into improving the reliability and accuracy of the XAI system for sentiment analysis in a low-resource language context.

5.5.4. Interpretability

Interpretability is defined as the ability to explain to a human being the decision made by an artificial intelligence model [5]. In this case, we further evaluated interpretability in the XAI methods by asking 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. *Were there any challenges or difficulties you encountered while trying to interpret the XAI system's explanations of sentiment?*

In this section, our participants were asked to answer the three questions related to interpretability. Figure 16 shows the results obtained from the participants.

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 suggest that the explanations offered by the XAI system significantly contributed to enhancing the participants' comprehension 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 also 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 sentiment 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 further show that the majority, which constitutes 79%, did not encounter challenges when interpreting the explanations provided by the XAI system regarding sentiment analysis. In contrast, a smaller percentage, accounting for 21%, acknowledged facing difficulties or obstacles when trying to understand the explanations of the XAI system. This distribution highlights a predominant ease in comprehending the system's provided explanations for sentiment analysis, although a notable minority found certain aspects challenging, signifying a need for potential improvements in explanation clarity or user guidance for enhanced comprehension.

6. Conclusions

In this paper, we propose an approach that combines XAI methods and Afro-centric models for sentiment analysis. Although sentiment analysis tools and techniques are available in English, it is essential to cover other low-resourced African languages, and the models make 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 suitable for transformer-based Afrocentric models. Although the models perform relatively well across the languages, the Afro-XLMR model outperformed all the models in this sentiment analysis task, showing its adaptability to low-resource languages. Furthermore, we used XAI tools to visualize the attention in the texts. Then, we 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, the 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 feedback survey additionally validate that a significant majority of the participants agree 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 the future, we are going to investigate the use of other XAI methods on these pre-trained models. We also plan to address problems of the perturbation-based explanation process when dealing with textual data. Furthermore, we intend to investigate counterfactual visualizations, in which we alter various elements to assess the impact of the explanations. Additionally, 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. However, it is important to acknowledge the limitations of this study, particularly in handling sarcastic phrases and idiomatic expressions that may lead to incorrect sentiment results in tweets processed by MPLM models.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Our dataset will be made publicly available and accessible by all NLP communities at https://github.com/NLPforLRLsProjects/SAfriSenti-Corpus.

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.

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

References

1. Vinuesa, Ricardo and Azizpour, Hossein and Leite, Iolanda and Balaam, Madeline and Dignum, Virginia, and Domisch, Sami and Felländer, Anna and Langhans, Simone Daniela and Tegmark, Max, and Fuso Nerini, Francesco. The Role of Artificial Intelligence in Achieving the Sustainable Development Goals. *Nature communications* **2020**, *11*, 1–10.
2. Sharma, H.D.; Goyal, P. An Analysis of Sentiment: Methods, Applications, and Challenges. *Engineering Proceedings* **2023**, *59*. doi:10.3390/engproc2023059068.
3. Kokalj, Enja and Škrlj, Blaž and Lavrač, Nada and Pollak, Senja and Robnik-Šikonja, Marko". BERT meets Shapley: Extending SHAP Explanations to Transformer-based Classifiers. Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Association for Computational Linguistics, 2021, pp. 16–21.
4. Fantozzi, P.; Naldi, M. The Explainability of Transformers: Current Status and Directions. *Computers* **2024**, *13*. doi:10.3390/computers13040092.
5. Alejandro Barredo Arrieta and Natalia Díaz-Rodríguez and Javier Del Ser and Adrien Bennetot and Siham Tabik and Alberto Barbado and Salvador Garcia and Sergio Gil-Lopez and Daniel Molina and Richard Benjamins and Raja Chatila and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion* **2020**, *58*, 82–115.
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. *Information Fusion* **2023**, *99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805.
7. Omnia Amin and Blair Brown and Bruce Stephen and Stephen McArthur. A case-study led investigation of explainable AI (XAI) to support deployment of prognostics in industry. Proceedings of the European Conference Of The PHM Society 2022; Do, P.; Michau, G.; Ezhilarasu, C., Eds., 2022, pp. 9–20.
8. United Nations. Sustainable Development Goals: 17 Goals to Transform our World. https://www.un.org/sustainabledevelopment/sustainabledevelopment-goals, 2022. Accessed: 2023-08.
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). *Computer Methods and Programs in Biomedicine* **2022**, *226*, 107161. doi:https://doi.org/10.1016/j.cmpb.2022.107161.
10. Kumar, P.; Hota, L.; Tikkiwal, V.A.; Kumar, A. Analysing Forecasting of Stock Prices: An Explainable AI Approach. *Procedia Computer Science* **2024**, *235*, 2009–2016. International Conference on Machine Learning and Data Engineering (ICMLDE 2023), doi:https://doi.org/10.1016/j.procs.2024.04.190.
11. Schoonderwoerd, Tjeerd A.J. and Wiard Jorritsma and Neerincx, Mark A. and Van Den Bosch, Karel. Human-centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *International Journal of Human-Computer Studies* **2021**, *154*, 1–25.
12. Minchae Song. A Study on Explainable Artificial Intelligence-based Sentimental Analysis System Model. International Journal of Internet, Broadcasting and Communication, 2022, pp. 142–151.
13. Wankhade, Mayur and Rao, Annavarapu and Kulkarni, Chaitanya. A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artificial Intelligence Review* **2022**, pp. 1–50.
14. Mabokela, Koena Ronny and Celik, Turgay and Raborife, Mpho. Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access* **2023**, *11*, 15996–16020.
15. Le, Tuan Anh and Moeljadi, David and Miura, Yasuhide and Ohkuma, Tomoko. Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. Proceedings of the 12th Workshop on Asian Language Resources (ALR12). The COLING 2016 Organizing Committee, 2016, pp. 123–131.
16. Adebara, Ife and Elmadany, AbdelRahim and Abdul-Mageed, Muhammad and Alcoba Inciarte, Alcides. SERENGETI: Massively Multilingual Language Models for Africa. Findings of the Association for Computational Linguistics: ACL 2023; Association for Computational Linguistics: Toronto, Canada, 2023; pp. 1498–1537.
17. Alabi, Jesujoba O. and Adelani, David Ifeoluwa and Mosbach, Marius and Klakow, Dietrich. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. Proceedings of the 29th International Conference on Computational Linguistics; International Committee on Computational Linguistics: Gyeongju, Republic of Korea, 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". "Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)"; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates (Hybrid), 2022; pp. 52–64.
19. Mabokela, Koena Ronny and Roborife, Mpho and Celik, Turguy. Investigating Sentiment-Bearing Words- and Emoji-based Distant Supervision Approaches for Sentiment Analysis. Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023); Association for Computational Linguistics: Dubrovnik, Croatia, 2023; pp. 115–125.
20. Marvin M. Aguero-Torales and Jose I. Abreu Salas and Antonio G. Lopez-Herrera. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing* **2021**, *107*, 107–373.
21. Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186.
22. Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* **2019**, *abs/1907.11692*.
23. Conneau, Alexis and Khandelwal, Kartikay and Goyal, Naman and Chaudhary, Vishrav and Wenzek, Guillaume and Guzmán, Francisco and Grave, Edouard and Ott, Myle and Zettlemoyer, Luke and Stoyanov, Veselin. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020, pp. 8440–8451.
24. Ogueji, Kelechi and Zhu, Yuxin and Lin, Jimmy. Small Data? No Problem! Exploring the Viability of Pre-trained Multilingual Language Models for Low-resourced Languages. Proceedings of the 1st Workshop on Multilingual Representation Learning; Association for Computational Linguistics: Punta Cana, Dominican Republic, 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*. doi:10.3390/electronics10182195.
26. Sage Kelly and Sherrie-Anne Kaye and Oscar Oviedo-Trespalacios. What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics* **2023**, *77*, 101925.
27. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* **2023**, *263*, 110273.
28. Qian, Kun and Danilevsky, Marina and Katsis, Yannis and Kawas, Ban and Oduor, Erick and Popa, Lucian and Li, Yunyao. XNLP: A Living Survey for XAI Research in Natural Language Processing. 26th International Conference on Intelligent User Interfaces - Companion; Association for Computing Machinery: New York, NY, USA, 2021; IUI '21 Companion, p. 78–80.
29. Liu, Shengzhong and Le, Franck and Chakraborty, Supriyo and Abdelzaher, Tarek. On Exploring Attention-based Explanation for Transformer Models in Text Classification. IEEE International Conference on Big Data, 2021, pp. 1193–1203.
30. Park, S.; Lee, J. LIME: Weakly-Supervised Text Classification without Seeds". Proceedings of the 29th International Conference on Computational Linguistics; International Committee on Computational Linguistics: Gyeongju, Republic of Korea, 2022; pp. 1083–1088.
31. Bodria, F.; Panisson, A.; Perotti, A.; Piaggesi, S. Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis. Sistemi Evoluti per Basi di Dati, 2020.
32. Marco Tulio Ribeiro and Sameer Singh and Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**.
33. Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *ArXiv* **2017**, *abs/1705.07874*.

34. Diwali, Arwa and Saeedi, Kawther and Dashtipour, Kia and Gogate, Mandar and Cambria, Erik and Hussain, Amir. Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis. *IEEE Transactions on Affective Computing* **2023**, pp. 1–12.
35. Mabokela, Ronny and Schlippe, Tim. A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context. Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages; European Language Resources Association: Marseille, France, 2022; pp. 70–77.
36. Muhammad, S.; Abdulmumin, I.; Ayele, A.; Ousidhoum, N.; Adelani, D.; Yimam, S.; Ahmad, I.; Beloucif, M.; Mohammad, S.; Ruder, S.; Hourrane, O.; Jorge, A.; Brazdil, P.; Ali, F.; David, D.; Osei, S.; Shehu-Bello, B.; Lawan, F.; Gwadabe, T.; Rutunda, S.; Belay, T.; Messelle, W.; Balcha, H.; Chala, S.; Gebremichael, H.; Opoku, B.; Arthur, S. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Bouamor, H.; Pino, J.; Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 13968–13981. doi:10.18653/v1/2023.emnlp-main.862.
37. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR, 2020.
38. Arras, L.; Montavon, G.; Müller, K.R.; Samek, W. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, 2017, pp. 159–168.
39. Vig, Jesse. A Multiscale Visualization of Attention in the Transformer Model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; Association for Computational Linguistics": Florence, Italy, 2019; pp. 37–42.
40. 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. https://doi.org/10.1016/j.neucom.2022.09.129.
41. Mabokela, Koena Ronny and Schlippe, Tim. AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa. Artificial Intelligence Research. Springer Nature Switzerland, 2022, pp. 309–322.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.