SAINT: Multilingual Span-Level Interpretability for Sentiment Analysis

Seid Muhie Yimam[1], **Tadesse Destaw Belay**[2], **Robert Geislinger**[1], **Shamsuddeen Hassan Muhammad**[3], **Adaeze Ngozi Ohuoba**[4], **Sukairaj Hafiz Imam**[5], **Abinew Ali Ayele**[6], **Martin Semmann**[1], **Chris Biemann**[1] **and Serge Sharoff**[4]
[1]University of Hamburg, Hamburg, Germany, [2]Instituto Politécnico Nacional, Mexico, [3]Imperial College London, London, UK, [4]University of Leeds, West Yorkshire, England, [5]Bayero University, Kano, Nigeria, [3]Bahir Dar University, Bahir Dar, Ethiopia,
{seid.muhie.yimam,robert.geislinger,martin.semmann,chris.biemann}@uni-hamburg.de
tadesseit@gmail.com, shamsuddeen2004@gmail.com,
{mlano, s.sharoff}@leeds.ac.uk, sukhimam00@gmail.com

Abstract

We investigate multilingual sentiment analysis and interpretability across high- and low-resource languages, focusing on Amharic, English, German, and Hausa. Our study evaluates encoder-only transformer models for both sequence-level sentiment classification and token-level attribution using Captum. Additionally, we assess zero- and few-shot decoder-only models for sequence-level sentiment prediction. Our results show that few-shot decoder-only models outperform encoder-only models on token-level sentiment classification in most languages, with the exception of Hausa, where a multilingual encoder-based model leads. For sequence-level sentiment classification, encoder-only models generally achieve strong performance across most languages, but decoder-only models are highly competitive, and may even surpass encoders, in the high-resource settings (German, English) and low-resource scenarios, depending on the prompting strategy. These findings highlight the utility of combining fine-tuned transformer models with prompt-based large language models to build interpretable sentiment analysis systems across both low- and high-resource languages. The SAINT dataset, annotation guideline, and evaluation scripts can be found at https://github.com/uhh-hcds/SAINT.

1. Introduction

Sentiment analysis represents a core topic within Natural Language Processing (NLP) that identifies and classifies opinions and emotional content expressed in textual data. The proliferation of user-generated content on platforms such as social media, customer reviews, and digital communications has positioned sentiment analysis as an essential analytical tool. Sentiment analysis now serves critical functions across business intelligence, public opinion research, and social science research, providing insights that connect textual data to human emotional responses in structured, actionable formats (Alam et al., 2025).

As sentiment analysis models are increasingly deployed in real-world applications, it is crucial not only to achieve high predictive accuracy but also to ensure that their predictions are interpretable and transparent. Interpretability enables users to understand and trust how these models arrive at their decisions, particularly in sensitive domains. Although recent developments in transformer-based models have significantly enhanced sentiment classification in high-resource languages such as English and German, many low- and medium-resource languages remain underrepresented. Consequently, particularly from an interpretability standpoint, there is an increasing need to evaluate how sentiment analysis models perform across both high-resource and low-resource language settings (Kokhlikyan et al., 2020; Ribeiro et al., 2016; Miglani et al., 2023).

Ensuring that sentiment analysis models provide comprehensible and transparent predictions is especially important for low-resource languages, where limited linguistic infrastructure can obscure biases and systematic errors (Dossou et al., 2022). When users understand the reasons behind a model's sentiment classifications, whether neutral, negative, or positive,

they are better equipped to validate results and trust model behavior, particularly in domains such as customer service, social media monitoring, political opinion analysis, and so on (Ribeiro et al., 2016; Kokhlikyan et al., 2020).

In this study, we focus on sentiment explainability across diverse language settings, evaluating four languages with different resource levels: **English** and **German** (high-resource), **Hausa** (medium-/low-resource), and **Amharic** (low-resource). We assess both **encoder-only fine-tuned transformer models** (BERT-base, BERT-base-german (Devlin et al., 2019), XLM-RoBERTa-base (Conneau et al., 2020), AfroXLMR-large (Dossou et al., 2022), and AfroXLMR-social (Belay et al., 2025)) as well as **decoder-only large language models** in a zero- and few-shot setting (GPT-4o-mini (OpenAI et al., 2024b), GPT-4.1-mini (OpenAI et al., 2025b), gpt-oss-120b (OpenAI et al., 2025a), Qwen3-32b (Team et al., 2025) and LlaMA-3.3-70B-it (Grattafiori et al., 2024; Meta, 2024)). We evaluate both **sequence-level sentiment classification** and **token-level sentiment classification** by fine-tuning encoder models, as well as conducting zero- and few-shot classification with LLMs. For interpretability, we employ the Captum (Miglani et al., 2023) library to compute token-level attributions for sentiment predictions, providing insight into which input tokens most strongly influence model decisions. This attribution analysis helps explain the reasoning behind model outputs (see Section 6.3 for details).

Our key contributions are as follows:

1. We conduct a comprehensive comparative analysis of sentiment explainability in English, German, Amharic, and Hausa using encoder-only fine-tuned models and decoder-only models with zero- and few-shot settings.

2. We implement sequence-level sentiment classification and fine-tune encoder-only models for token-level sentiment classification, and further provide token-level interpretation using Captum-based attribution methods.

3. We highlight challenges and differences in interpretability across languages, resource levels, and model types.

4. We have created an annotated multilingual sentiment explainability dataset, **SAINT**, to advance research in interpretability.

This work is guided by the following research questions:

1. **RQ1**: How do transformer models differ in their ability to explain sentiment predictions in high, medium, and low-resource languages?

2. **RQ2**: Do decoder-only generative models provide more human-understandable explanations than encoder-only attribution methods?

3. **RQ3**: What are the challenges of achieving reliable interpretability for both encoder and decoder models in high- and low-resource language settings?

Through these investigations, we aim to provide an in-depth understanding of explainable sentiment analysis across linguistically diverse settings.

Note that in this work, we use the term *span-level* and *token-level* interchangeably, where both refer to a natural language word-level annotation, which is different from subword or subtoken units known in the transformer-based tokenizer.

2. Related Work

Sentiment analysis has progressed significantly, transitioning from lexicon-based and classical machine learning approaches, such as SVMs and logistic regression, to deep learning architectures like CNNs to Transformer based approaches. However, these earlier models struggled with capturing long-range dependencies and contextual semantics in unstructured, multilingual data (Geislinger, 2024).

The advent of transformer-based models like BERT (Devlin et al., 2019) has greatly improved performance across NLP tasks by enabling better context-aware representations. Multilingual models such as mBERT and XLM-R (Conneau et al., 2020) extended this capability to multiple languages, enabling transfer learning, with continual training or finetuning (Belay et al., 2025), for low-resource settings.

AfroXLMR, a variant of XLM-R, specifically fine-tuned for African languages including Amharic and Hausa (Dossou et al., 2022), represents an important step toward inclusion of under-represented languages in NLP research.

While significant work exists in multilingual sentiment classification, much less attention has been given to the *interpretability* of these models, especially across languages with different resource levels. Most prior work either focuses on model performance or uses monolingual interpretability approaches, which do not generalize well to cross-lingual or token-level sentiment insights.

Interpretability methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) have been widely used, but are often computationally expensive and model-agnostic. In contrast, the Captum library (Kokhlikyan et al., 2020) offers efficient, model-specific explainability techniques, which are well-suited for transformer models.

In parallel, the emergence of large-scale *decoder-only language models* such as GPT (OpenAI et al., 2024a), Llama (Grattafiori et al., 2024), gpt-oss (OpenAI et al., 2025a), and Qwen (Team et al., 2025) has introduced new possibilities for multilingual sentiment analysis in zero-shot or few-shot settings. These models can not only perform classification without explicit fine-tuning but also generate natural language rationales for predictions. However, their performance in resource-diverse languages such as Amharic, English, German, and Hausa, and their ability to provide reliable explanations, remain underexplored.

3. Data Collection and Annotation

Annotating sentiment analysis data with explainability, here identifying not only the overall sentiment but also the specific positive and negative spans, and their associated entities, is a highly complex and time-consuming task. To ensure high-quality and consistent annotations, we deliberately avoided the traditional crowdsourcing approach, which is often feasible for standard sentiment tasks but less reliable for fine-grained, explainable annotations. Instead, we employed one trained expert annotator per language. Rather than compensating on a per-instance basis, we hired these expert annotators for a period of three months, during which they participated in data collection, adapted the Potato annotation tool for our needs (see Figure 1), and carried out the detailed annotations (Pei et al., 2022). Throughout the process, more than three co-authors of this paper were actively involved in overseeing, reviewing, and validating the annotations to ensure accuracy and consistency. Hence, as is common practice in expert-driven annotation workflows for complex tasks, we do not report inter-annotator agreement, as we relied on a trained annotator and consistently checked the quality by the co-authors.

All annotations are performed at both the sequence level (labeling overall sentiment as Positive, Negative, or Neutral) and the token/span level, where specific sentiment expressions and their contextual targets or objects (such as named entities) were annotated for fine-grained analysis.

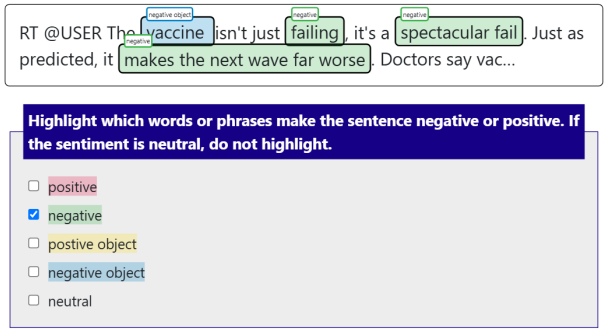

Figure 1: English Sentiment Interpretability Annotation task example

3.1. Amharic Dataset

The Amharic dataset comprises 1,519 social media and news comments collected from X (formerly Twitter; we use data from before 2023, when Twitter data was allowed for research), and regional news platforms. These posts reflect a broad spectrum of public opinion on political developments, social tensions, and cultural narratives within Ethiopian discourse.

3.2. English Dataset

The English dataset consists of 1,501 social media posts selected from our study on anti-vaccination discourse. Unlike the other

dataset, these posts mostly focus on one topic, but they reflect a broad spectrum of functional styles, including texts aimed at expressing opinions, sharing stories, or promoting products and services.

3.3. German Dataset

For the German language, we curated a dataset of 1,498 social media posts sourced from publicly available social media platforms such as X(Twitter) and Reddit. These posts cover a wide range of topics, including geopolitics, social debates, and cultural commentary, ensuring a diverse and context-rich representation of public sentiment.

3.4. Hausa Dataset

The Hausa dataset consists of 1,879 comments sourced from online forums, regional news portals, and social media platforms. The dataset captures sentiments on the political, religious, and cultural identity domains where Hausa is widely used throughout West Africa.

The summarized data statistics with 60:10:30 (train, dev, and test, respectively) splitting ratio are shown in Table 1, and the detailed class-level sentiment distributions are shown in Figure 2.

Language	Train	Dev	Test	Total
Amharic (amh)	912	153	454	1,519
English (eng)	899	151	451	1,501
German (deu)	899	150	449	1,498
Hausa (hau)	1,127	189	563	1,879

Table 1: Dataset statistics: number of annotated sentences across languages.

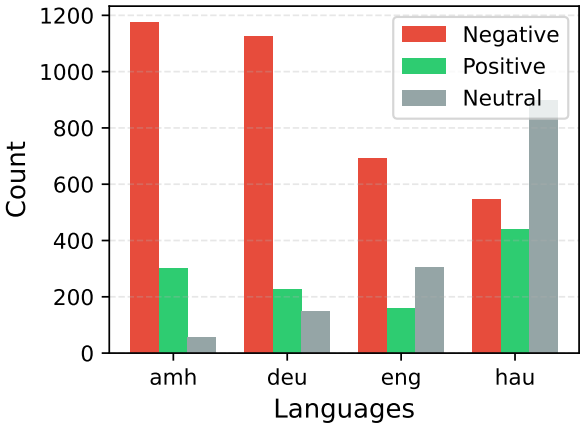

Figure 2: Data statistics across languages and sentiment labels.

4. Experimental Setup

We employ a combination of encoder-only transformer models and decoder-only large language models (LLMs) to study multilingual sentiment explainability. This dual setup enables us to evaluate both traditional supervised fine-tuning and prompt-based inference.

Encoder-only Transformers: For English we use Roberta-base and XLM-Roberta-base models. For German, we use BERT-base-german, a monolingual BERT variant pretrained on large German corpora. For Amharic and Hausa, we employ AfroXLMR (Dossou et al., 2022) and AfroXLMR-Social (Belay et al., 2025), a multilingual transformer specifically fine-tuned for African languages. Both models are fine-tuned on our annotated datasets for sequence-level and token-level sentiment classification. We also use the Captum library (Kokhlikyan et al., 2020) to provide gradient-based interpretability for these transformer models.

Decoder-only LLMs: We also evaluate generative models such as GPT-4.1-mini, GPT-4o, LLaMA 3.3, gpt-oss, and Qwen3 in a zero- and few-shot setting. These models are not fine-tuned on our datasets; instead, handcrafted prompts are designed to elicit sentiment predictions at both the sequence and token levels.

Finetuning Strategy: For the encoder-only models, we employed both monolingual and multilingual fine-tuning approaches. In monolingual fine-tuning, each model is trained separately on the training split of its target language, as indicated in Table 1. For multilingual fine-tuning, the training and development splits from all languages are merged to create a larger and more diverse training set. Models fine-tuned under this setting are exposed to linguistic variety during training, but evaluation is still performed on the original test set for each language to ensure fair comparisons. This dual approach allows us to assess the impact of monolingual versus multilingual train-

ing on both overall performance and model generalization.

Comparative Setup: This architecture choice allows us to contrast supervised fine-tuning with encoder-only transformers[1] against prompt-based inference using decoder-only LLMs. While transformers are efficient for structured classification and feature-level attribution, LLMs offer flexible interpretability through explanations and reasoning steps. Together, these complementary approaches highlight interpretability gaps across architectures, languages, and resource levels.

5. Task Formulations

Sequence Classification: The goal is to classify the overall sentiment of a given text as Positive, Negative, or Neutral. In this task, the model takes the entire sentence or post and predicts a single label. Sequence-level classification provides a high-level overview of sentiment trends, which is useful for applications such as social media monitoring, customer feedback analysis, and public opinion tracking. By evaluating sequence classification across multiple languages, we can assess how well models generalize from high-resource languages (English and German in our case) to low- and medium-resource languages (Amharic and Hausa), and whether interpretability methods, specifically Captum library attribution scores, highlight language-specific biases or errors.

Token Classification: Token-level sentiment analysis requires the identification of individual words or phrases that contribute to the sentiment of the text. We use the BIO (Begin, Inside, Outside) tagging scheme to prepare the tokens for classification as part of positive, negative, or neutral sentiment spans. This fine-grained approach enables span-level interpretability, allowing us to pinpoint which parts of a sentence are driving the model's prediction. Token classification also facilitates error analysis and comparison between encoder-based transformers and decoder-only LLMs in their ability to highlight sentiment-bearing tokens. Moreover, this task is particularly challenging for morphologically rich or low-resource languages, where sentiment expressions may be implicit, compound, or context-dependent. Through token-level evaluation, we aim to understand how well models capture nuanced sentiment patterns and provide comprehensible explanations.

Evaluation Metrics: For sequence-classification sentiment analysis, we used the weighted average F1 score. For token-level prediction, we report the CoNLL F1 score at the token-level, computed using the conlleval script[2].As token-level accuracy is often considered less informative for phrase-level tasks like NER or chunking, as it can be inflated by correctly identified "O" (outside of a chunk) tokens, we used F1 scores from conlleval. Conlleval metrics ignore others ("O") and focus on the two classes of tokens, "B-positive" and "B-negative."

6. Results

We evaluated both encoder-only models with task-specific fine-tuning and decoder-only models in a zero- and few-shot setting across Amharic, English, German, and Hausa datasets. Table 2 summarizes the F1-scores for both sequence-level sentiment classification and token-level classification tasks.

6.1. Encoder-only Models

For sequence-level sentiment classification, **AfroXLMR-Social** achieved the best overall performance, with F1 scores of 89.84/90.78(mono-/multi-lingual finetuning) on Amharic, 69.93/54.90 on English, 54.88/71.31 on German, and 92.86/92.97 on Hausa.

For the token-level classification task, overall performance was lower compared to sequence-level classification, reflecting the increased difficulty of precisely identifying sentiment-bearing spans. However, **AfroXLMR-large** and **XLM-RoBERTa-base**

[1] Encoder-only models are fine-tuned for both sentiment and token-level classification with the following hyperparameters: epochs 5 and batch size of 16. The detailed LLM prompts will be made available in the camera-ready version, along with GitHub links to the dataset.

[2] https://github.com/sighsmile/conlleval

achieved the strongest results, particularly on Hausa and Amharic.

6.2. Decoder-only Models

Among the zero-shot decoder-only LLMs, **GPT-4o-mini** produced the best results for sequence-level sentiment classification, with an average F1 of 74.14, excelling on Amharic (76.94) and Hausa (66.67), showing strong scores on both German (80.89) and English (72.06). **GPT-4.1-mini** also performed competitively, achieving an average of 68.30. The open weight model **gpt-oss-120b** performed competitively on Amharic, English, and German, while underperforming for Hausa. **Llama-3.3-70b** performed best on German while underperforming on English and Hausa. In general, **Qwen3-32b** struggled, particularly on Amharic, Hausa, and even English, resulting in an average of 56.42.

Among the few-shot decoder-only LLMs, **GPT-4o-mini** produced the best results for sequence-level sentiment classification, with an average F1 of 74.74, excelling on Amharic (79.82) and English (72.73) and showing strong scores on both German (80.00) and Hausa (66.67). **GPT-4.1-mini** and **Llama-3.3-70b** also performed competitively, achieving averages of 72.16 and 70.46, respectively. In contrast, **Qwen3-32b** struggled, particularly on Amharic and English, while performing the best 3-shot result for Hausa, resulting in an average of 63.67.

For the token-level task, performance was generally lower across all decoder-only models, but most performed better than the encoder-only models. **GPT-4.1-mini** achieved the best results on most languages with an average of 37.00 for zero-shot and 39.69 for few-shot. It achieved the best results for Amharic, German, and English. Only for Hausa **GPT-4o-mini** zero-shot achieved better results, but had only an average of 31.29.

Qwen3-32b achieved the lowest performance for both zero- and few-shot, but improved the average the most, from 19.29 at least up to 28.78, which is an improvement of 49%. The differences for all other models are mixed. While **Llama-3.3-70b** gains 35% for English with few shots compared to zero-shot, it loses 5% for Hausa. In general, nine few-shot experiments showed significantly better results, five equal, and six lower results, showing the general possibilities of In-context learning. The results are shown in Table 3.

Models	amh	eng	deu	hau
Encoder only models - Monolingual finetuning				
BERT-base	67.72	66.76	**57.46**	89.16
BERT-base-deu	67.81	64.66	52.71	87.65
RoBERTa-base	67.72	66.76	**57.46**	89.16
XLMR-base	87.56	64.66	57.39	87.40
AfroXLMR-large-76L	87.03	64.66	50.00	89.90
AfroXLMR-Social	**89.84**	**69.93**	54.88	**92.86**
Encoder only models - Multilingual finetuning				
BERT-base	67.81	59.10	64.66	89.18
BERT-base-deu	67.81	50.77	65.18	86.18
RoBERTa-base	68.32	59.21	65.15	88.30
XLMR-base	86.55	**60.91**	71.28	89.37
AfroXLMR-large-76L	83.93	50.62	64.66	66.97
AfroXLMR-Social	**90.78**	54.90	**71.31**	**92.97**
LLM results (Zero / three shot)				
GPT-4o-mini	76.94	72.06	80.89	**66.67**
GPT-4o-mini(3-shot)	**79.82**	**72.73**	80.00	66.41
GPT-4.1-mini	71.40	71.84	78.44	51.52
GPT-4.1-mini(3-shot)	75.39	71.18	78.67	63.38
gpt-oss-120b	73.17	72.50	80.67	45.96
gpt-oss-120b(3-shot)	75.17	70.95	79.56	54.29
Llama-3.3-70b	71.75	52.67	**80.98**	52.78
Llama-3.3-70b(3-shot)	78.30	69.84	80.44	53.28
Qwen3-32b	43.79	65.65	75.29	40.96
Qwen3-32b(3-shot)	44.39	67.05	75.85	66.58

Table 2: **Sentiment classification results** (weighted F1). All open source models are instruction-tuned versions.

6.3. Explaining Sentiment Predictions via Captum Library Integrated Gradients

To provide interpretability and transparency for transformer-based sentiment classification models, we employ the Captum library's Integrated Gradients method to estimate the contribution of each input word to the model's sentiment prediction. This involves applying random infinitesimal perturbations to the input embeddings to trace the gradients of these changes through the model. Specifically, for each test sample, we compute token-level attributions by applying Integrated Gradients on the input embedding layer of each fine-tuned model. The resulting attributions quantify the relative importance of each word towards the model's predicted sentiment class.

We aggregate subword attributions to word-level values by summing over all subwords

Model	amh	eng	deu	hau
Encoder only models Monolingual / Multilingual				
BERT-base	02.30 / 05.45	**10.56** / 13.51	06.91 / 10.61	38.84 / 38.91
BERT-base-deu	01.13 / 00.00	04.28 / 01.18	19.01 / 13.82	11.42 / 22.72
RoBERTa-base	08.75 / 08.76	06.39 / 15.75	13.56 / 19.66	35.22 / 27.28
XLM-RoBERTa-base	31.30 / 31.51	08.05 / 11.81	24.47 / 23.34	**39.46** / **39.54**
AfroXLMR-large-76L	**33.13** / **31.94**	00.00 / **15.78**	**28.89** / **27.07**	26.31 / 34.38
AfroXLMR-Social	31.54 / 31.79	07.96 / 15.64	15.32 / 25.24	35.56 / 35.77
Zero / three shot for decoder-only				
GPT-4o-mini	30.86 / 28.83	26.90 / 27.05	30.29 / 29.41	**37.11** / 36.63
GPT-4.1-mini	**41.36** / **43.88**	**34.79** / **35.21**	**38.21** / **42.52**	33.63 / **37.13**
Llama-3.3-70b	32.11 / 32.75	21.09 / 28.51	35.96 / 34.18	34.01 / 32.44
gpt-oss-120b	33.20 / 34.85	32.39 / 31.91	31.66 / 32.35	22.38 / 30.20
Qwen3-32b	15.61 / 21.12	27.58 / 31.46	24.72 / 35.71	09.23 / 26.82

Table 3: **Span-level sentiment analysis** (F1 score) results across languages and models. The first row is span-level monolingual vs multilingual results from encoder-only models. The second row is the results from LLMs with zero and a few shots (3-shot).

corresponding to each whitespace-delimited token, and we discretize these into token-level sentiment labels ("positive", "negative", "neutral") using a threshold (set to 0.0 in this study). To evaluate the correspondence between model explanations and human interpretability, we compare the tokens with the largest positive or negative attributions to human-annotated sentiment span tags. This comparison yields token-level F1 scores (denoted FB1, computed with the Conll NER evaluation script), which are reported in Table 4.

Across all languages and models, we observe that the FB1 scores for token-level agreement between Captum-based attributions and human span annotations are low, most often in the 5–25% range (see Table 4). Several factors explain this relatively weak overlap between the words highlighted by the model and those annotated by humans:

- **Span Consistency vs. Token Attribution Mixing:** Our annotation guidelines require labeling contiguous spans as either positive or negative, avoiding mixed polarity within the same span. In contrast, Captum assigns each token an attribution score reflecting its contribution towards either class, potentially leading to both positive and negative attributions within the same span. When these token-level attributions are converted into categorical sentiment labels, this mixing of polarities increases disagreement with the human gold labels and thus lowers the overlap scores.

- **Captum Library Attributions Reflect Model, Not Human, Reasoning:** Integrated Gradients reveals what the model, in its own logic, 'uses' to make a prediction, examples shown in Appending A.

- **Tokenization and Alignment Issues:** Discrepancies between how words are split for gold annotation and how subwords are merged for Captum library scores can cause mismatches.

- **Impact of Error Propagation from Sentence-Level Misclassification:** The actual performance of the sentiment classifiers is also low (see Table 2); for some languages, models are achieving as low as 50% F1. If the model incorrectly predicts the overall sentence sentiment, its word-level attributions will almost always fail to align with the human-annotated sentiment spans (since it "rationalizes" its own misprediction). Thus, the span-level agreement will be significantly lower than sentence classification performance due to error propagation from misclassified in-

stances.

- **Limited Training Data:** The performance of both the sentence-level classifiers and the quality of attributions may be constrained by the small size of the training datasets (1k). With more training data, the models could learn more reliable alignment between explicit sentiment cues and predictions, potentially improving both classification accuracy and token-level explanation quality.

The low F1 scores emphasize that post-finetuning interpretability methods like Captum, while useful to understand neural network attention, do not always align with the explicit triggers that humans use.

Model	amh	deu	eng	hau
BERT-base	12.58	20.24	06.78	13.60
BERT-base-deu	**23.28**	12.93	08.74	14.63
RoBERTa-base	08.17	10.04	06.05	10.69
XLMR-base	16.14	18.23	07.16	**16.64**
AfroXLMR-large-76L	13.44	22.40	**13.24**	15.05
AfroXLMR-Social	19.26	**23.53**	08.71	15.42

Table 4: Token-level F1 agreement between Captum Integrated Gradients explanations and human-annotated sentiment span labels across models and languages.

6.4. Comparative Analysis

Our results highlight important differences between encoder-only and decoder-only models for both sequence-level and token-level sentiment classification across languages.

For sequence-level classification, encoder-only models such as AfroXLMR-Social attain strong results across all languages. These models are especially competitive in low-resource languages like Amharic and Hausa. However, instruction-tuned decoder-only LLMs (e.g., GPT-4o-mini and LLaMA-3.3-70B), used in zero- and few-shot settings, are able to match, or in some cases surpass, encoder models in sequence-level F1, mostly on high-resource languages. This demonstrates the flexibility and cross-lingual adaptability of advanced LLMs, even in settings with limited labeled data.

For token-level (span-level) sentiment classification, our findings show a more nuanced pattern. Decoder-only models, in particular GPT-4.1-mini in the few-shot setting, achieve the strongest F1 scores on Amharic, English, and German, outperforming their encoder-based counterparts in these languages. In contrast, for Hausa, the highest performance is achieved by the encoder-only XLM-RoBERTa-base model with multilingual fine-tuning. We witnessed that instruction-tuned LLMs can provide better results for token-level sentiment predictions even in the absence of explicit supervised training. Table 4 indicates the disagreement between the two views, as our annotators focused on what was important for the problem.

7. Conclusion

In this work, we developed a comprehensive and interpretable multilingual sentiment analysis framework, leveraging both encoder-only and decoder-only language models for sequence-level and token-level (span-level) classification. To enable fine-grained, high-quality human annotations, we customized the Potato annotation tool (Pei et al., 2022) for the span-level sentiment task and relied on expert annotators to produce reliable data across Amharic, English, German, and Hausa.

For modeling, we explored both monolingual and multilingual fine-tuning strategies for encoder-based transformers, including AfroXLMR, XLM-RoBERTa, and BERT variants, and systematically benchmarked them against prompt-based decoder-only LLMs such as GPT-4o-mini and in zero-shot and few-shot settings. This dual approach allowed us to rigorously evaluate both sequence-level and token-level sentiment classification tasks.

For sequence-level sentiment classification, encoder-only models such as AfroXLMR-Social achieved strong performance across all languages. At the same time, decoder-only LLMs, such as GPT-4o-mini, matched or outperformed encoders in certain languages and few-shot scenarios, highlighting their impressive cross-lingual transfer and flexibility with limited supervision. On token-level sentiment classification, instruction-tuned LLMs such as GPT-4.1-mini led for Amharic, German, and

English, while multilingual encoder-only models like XLM-RoBERTa-base achieved the best results on Hausa. This suggests that prompt-based LLMs are powerful for human-readable explanation extraction.

A very important question for further research is the ability to use explainability tools, such as the Integrated Gradients, to distinguish between model-based and problem-based explanations (Herrera, 2025). The model-based explanations highlight which features are important for the model, thus they are of relevance for the NLP researchers. The attributions are usually at the token level. The problem-based explanations are the rationales for the users for making their choices, so they often go beyond individual tokens, for example, covering syntactic constructions or stylistic choices. Table 4 indicates the disagreement between the two views, as our annotators focused on what was important for the problem. Our next research goal is to investigate alignment between the model attributions and human-annotated rationales beyond the token level. Specifically for the LLMs, while they can produce explanations behind their reasoning decisions, the explanations do not often align with the actual distribution of their weights, highlighting the limits of their "understanding" (West et al., 2024): they might be faithful to the problem, but to the model. More research is needed to make their explanations more faithful on the model level.

Limitation

Our work is not without limitations. First, the dataset was annotated by a single expert annotator per language. While the annotators are native speakers of the target language and experts in the same research area, individual subjective biases might influence the "gold standard" labels, particularly for more nuanced span-level sentiment. Further work can add further annotations on top of our dataset using three or five more annotators, report inter-annotator agreement, and explore the individual annotation effects compared to our data. The dataset size is also relatively small (ranging from 1,498 to 1,879 examples per language) to train encoder-only PLMs. Second, our LLM selection and contextual learning are limited to a few LLMs and zero and three-shot learning. A more contextual learning sample, such as five to ten shows, might improve the understanding of such explainable tasks. Third, we evaluated the sentiment (sequence classification) and the explainable tasks (span-level classification) as standalone individual tasks. Integrating both sentiment and the span level tasks in a multitasking approach might improve the performance of the tasks; it is recommended to explore further.

8. Acknowledgement

The SAINT project has received financial support from both the University of Hamburg (UHH) and the University of Leeds (UoL). Moreover, we have used the Scientific Compute Cluster at GWDG, the joint data center of Max Planck Society for the Advancement of Science (MPG) and the University of Göttingen.

9. Bibliographical References

Md Shah Alam, Md Sabbir Hossain Mrida, and Md Atikur Rahman. 2025. Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). *American Journal of Scholarly Research and Innovation*, 4(01):63–100.

Tadesse Destaw Belay, Israel Abebe Azime, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Idris Abdulmumin, Abinew Ali Ayele, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. AfroXLMR-Social: Adapting Pre-trained Language Models for African Languages Social Media Text. *arXiv preprint arXiv:2503.18247*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 8440–8451, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. 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)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid).

Robert Georg Geislinger. 2024. Enhancing sentiment analysis: Model comparison, domain adaptation, and lexicon evolution in german data. Master's thesis, Universität Hamburg, Germany, 7.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Francisco Herrera. 2025. Making Sense of the Unsensible: Reflection, Survey, and Challenges for XAI in Large Language Models Toward Human-Centered AI. *arXiv preprint arXiv:2505.20305*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4768–4777, Long Beach, CA, USA.

Meta. 2024. Meta Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: 2025-10-22.

Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using Captum to Explain Generative Language Models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore.

OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2024a. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

OpenAI, Sandhini Agarwal, Lama Ahmad, and Jason Ai et al. 2025a. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*.

OpenAI, Aaron Hurst, Adam Lerer, and Adam P. Goucher et al. 2024b. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

OpenAI, Ananya Kumar, Juahui Yu, John Hallman, and Michelle Pokrass et al. 2025b. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-10-22.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*, KDD '16, pages 1135–1144, New York, NY, USA.

Qwen Team, An Yang, Anfeng Li, and Baosong Yang et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The Generative AI Paradox: "What It Can Create, It May Not Understand". In *International Conference on Representation Learning*, volume 2024, pages 47159–47187, Vienna, Austria.

Appendix

A. Explainability span-level examples using Captum library

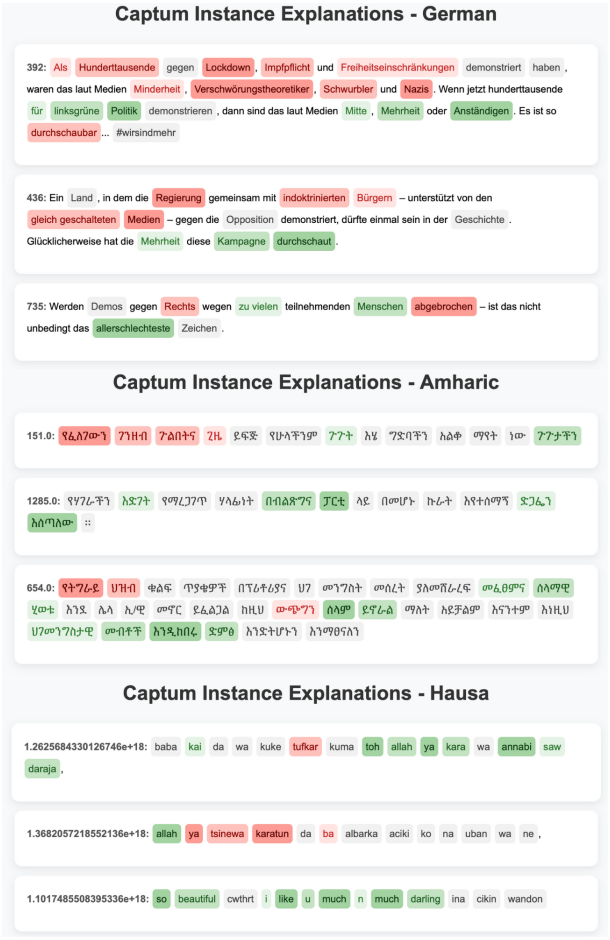

Figure 3: Explainable prediction examples using the Captum library from the XLM-RoBERTa-base model for German, Amharic, and Hausa language datasets.

B. LLM Prompt

Prompt 1

You are a sentiment and explainability annotation expert.
Given a text, perform the following steps:

1. **Assign Overall Sentiment:** Assign the overall sentiment for the text as one of: "positive", "negative", or "neutral".

2. **Explain the Sentiment:**
   - Briefly justify your sentiment label in one sentence.
   - Format: ‘"explanation": "<rationale>"‘

3. **Word-level Sentiment Tagging:** Split the text into a list of words using whitespace; keep punctuation attached to the word (e.g., "happy!"). - For each word:
   - Tag as "B-positive" if it expresses a positive sentiment and the overall sentiment is "positive".
   - Tag as "B-negative" if it expresses a negative sentiment and the overall sentiment is "negative".
   - Otherwise, tag as "O".
   - If the overall sentiment is "neutral", tag all words as "O".
   - The ‘words‘ and ‘tags‘ arrays must always be the same length.

4. **Output Format:** Return a single JSON object in the following format:
   {
   "sentiment": "<positive | negative | neutral>",
   "explanation": "<rationale>",
   "words": ["word1", "word2", ...],
   "tags": ["O", "B-positive", ...]
   }

   Key Points:
   - Do *not* tag words that express a different sentiment than the overall text.
   - If overall sentiment is "neutral", all tags should be "O".
   - Always keep the "words" and "tags" lists the same length.
   - For longer texts, apply these instructions to the whole text.