INF 791 Applied data sciece

Expanding a Multilingual Lexicon for Low-Resource Language Processing

Introduction

AI and Explainability

- Intelligent systems are reshaping society through machine learning (ML) innovations (Dwivedi et al., 2023).
- Explainable AI (XAI) helps users understand ML model decisions — essential for trust in critical sectors.

Lexicons and Low-Resource Languages

- Communication barriers across diverse African languages highlight the need for standardized sentiment lexicons (Lawless & Civille, 2013).
- Lexicon-based methods simplify sentiment interpretation but require extensive human input (Sadia et al., 2018).
- Many African languages are low-resource, with limited digital data and linguistic tools (Magueresse et al., 2020).

Lexicon Expansion for Low-Resource South African Languages

Original Lexicon

- 6,963 Ciluba and French words with sentiment scores (-9 to +9) and part of speech tags
- After cleaning duplicates (317), 6,646 unique entries remained
- Preprocessing steps included: stripping spaces and replacing multiple spaces with single space

Expansion Process

- Added columns for English, Zulu, Afrikaans, Sepedi, Xhosa, Shona
- Two-step translation pipeline: 1)French words were translated to English
  2)English words translated to the target South African languages
- Translation tool: Deep Translator (Google Translate API)
- Group members manually checked and fixed any translation errors

Challenges

- Google Translate struggles with low-resource languages (Maji et al., 2025)
- During the translation their were minor missing translations, due to API limitations, special characters or missing direct equivalents

Strengths: Process follows an efficient pipeline, covers multiple languages and involves manual corrections

Limitations: Minor translation gaps and heavy reliance on machine translation

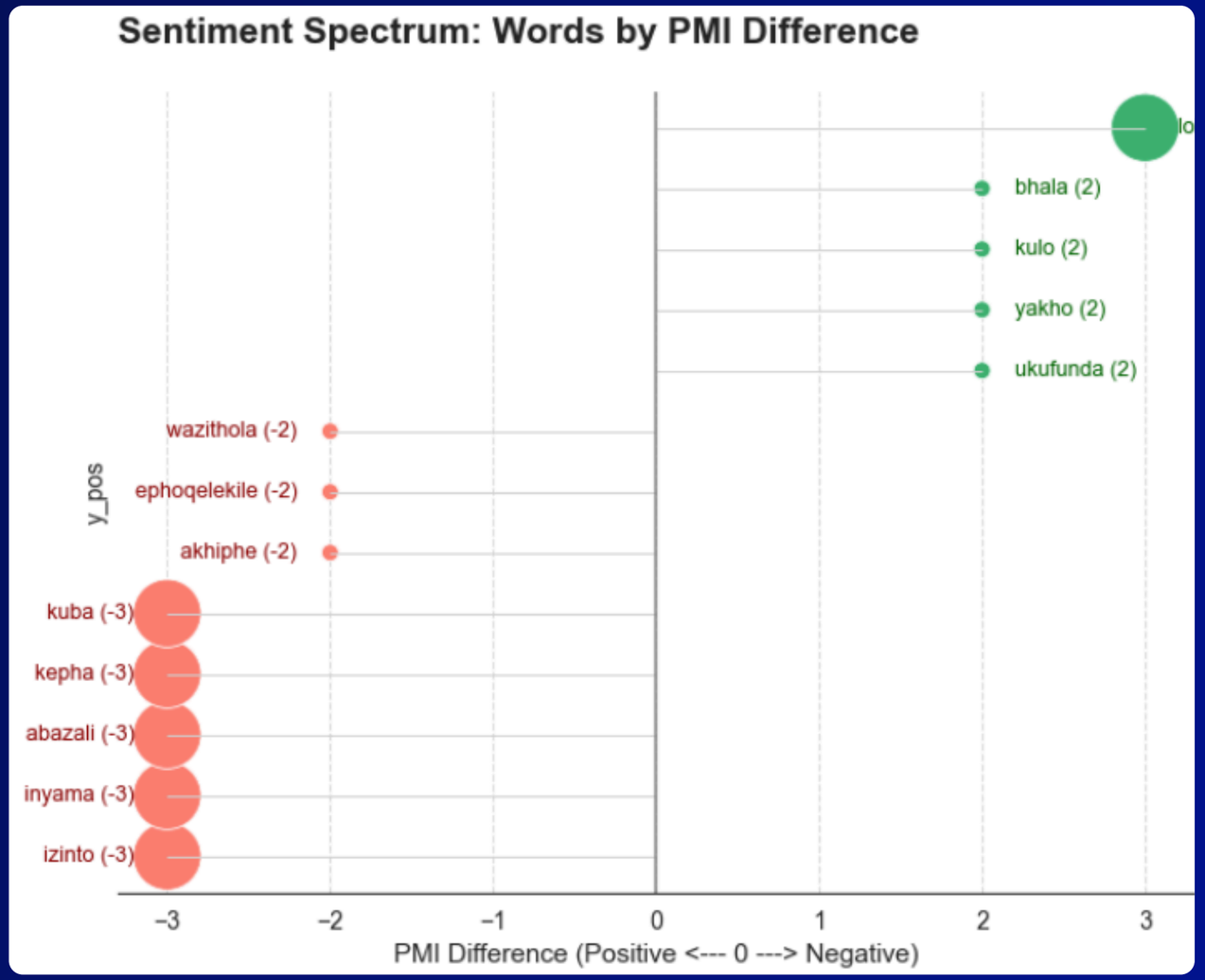
Text Normalization and Tokenization on Expanded Lexicon

Step	Description	Tools/Methods
Lowercasing	Standardise all words to lowercase	.lower() function
Spelling Correction	Hybrid approach: automatically translate + manual review corrections	PySpellChecker (English) RapidFuzz (Non-English)
Punctuation Removal	Removed commas, periods, exclamations. Kept Hyphens and apostrophes (carry linguistic meaning)	String libary
Whitespace & Special Character Cleaning	Removed tabs, excess space and non-alphanumerics	Expressions
Tokenization	For model building	Whitespace-based

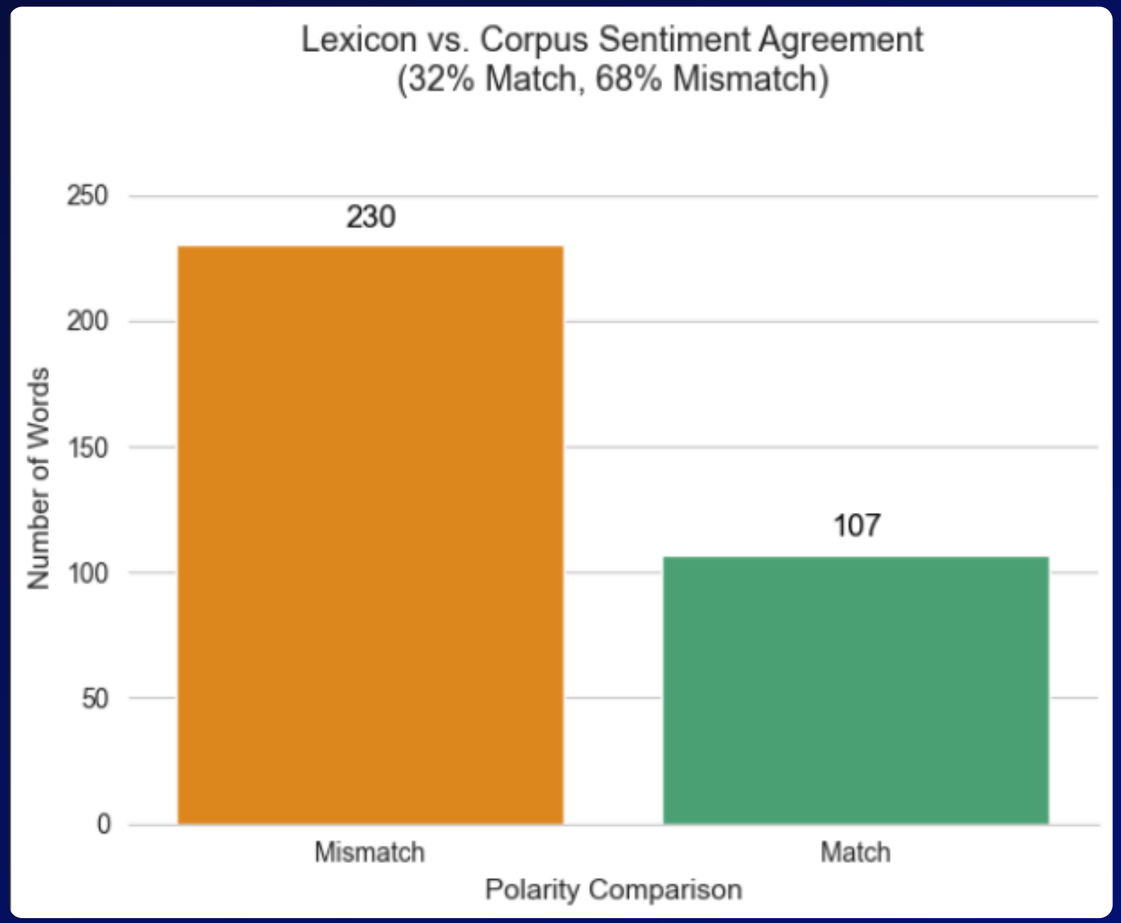
Corpus Based Enrichment

- We use real-world text (a corpus from the SADiLAR) to refine and expand lexicons sentiments.
- Compute Pointwise Mutual Information (PMI) for sentiment-bearing words.
- Polarity Assignment using PMI and Retrieval Augmented Generation (RAG).

1. **Tokenize Corpus:** Break text into words for analysis.
2. **Calculate PMI:** Measure association between words and sentiment words to calculate their polarity.
3. **Lexicon Comparison:** Compare corpus-inferred polarity to existing lexicon values to spot mismatches.
4. **Reassign Polarity:** For mismatched words, update the lexicon using corpus and external knowledge (like RAG models).



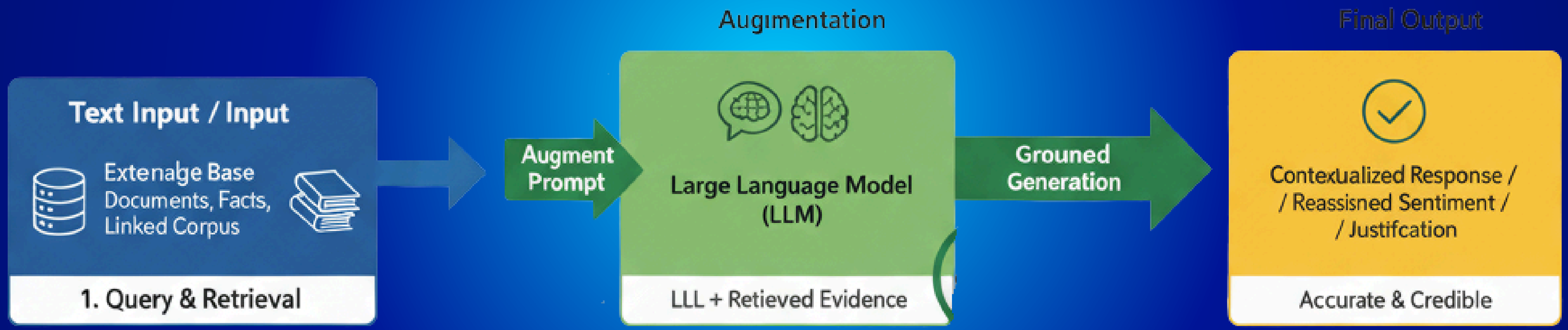
Polarity Mismatches



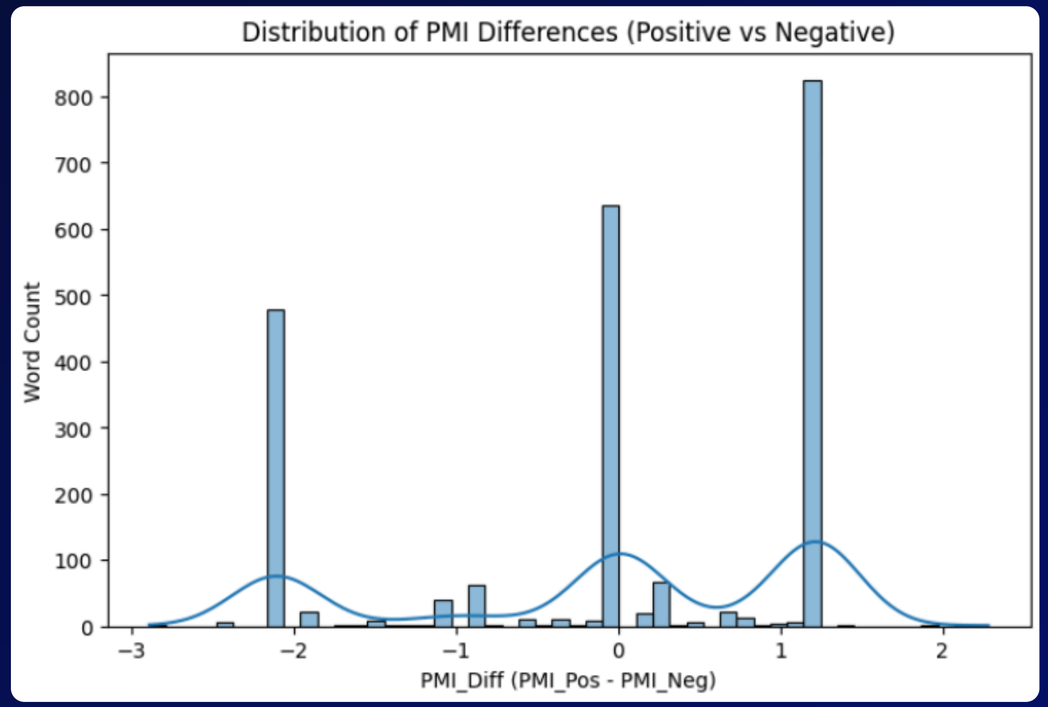
- Flagged 230 words where the strong corpus usage conflicts with the existing lexicon sentiment, highlighting key candidates for review

Polarity Mismatches Continued...

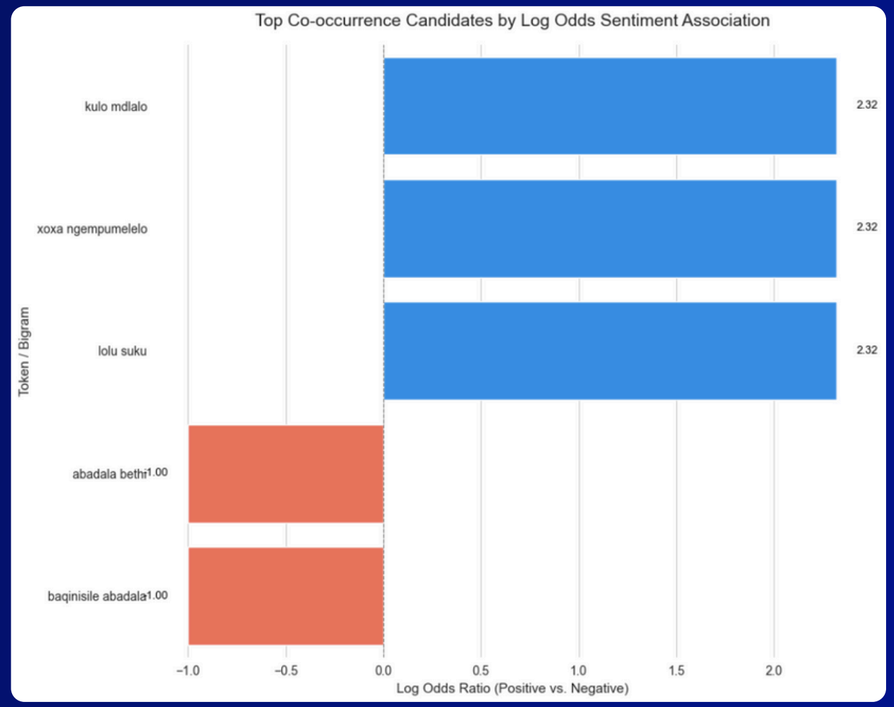
RAG helps by linking corpus findings with external databases for improved accuracy and credibility.



Statistical Analysis of Corpus Words



The corpus indicates a vocabulary that is not only highly polarized but also slightly biased towards positive language, as the highest frequency of words falls into the PMI_Diff > 1 range



Positive sentiment is far more defined by bigrams (kulo mdlalo, xoxa ngempumelelo, lolu suku), top pairs showing a high log odds ratio (+2.32), meaning they appear much more frequently in positive contexts than negative ones.

LIMITATIONS

- Need big, relevant corpora; rare words may be missed.
- Bias and ambiguity in the corpus can affect results.
- RAG adds power but also complexity and can introduce its own biases.

AfriBERTa and AfroXLMR

Purpose

- To understand the African-focused transformer models used for multilingual sentiment analysis.
- To highlight how each model addresses low-resource language challenges.

AfriBERTa

- Based on XLM-RoBERTa, fine-tuned for 11 African languages.
- Handles text classification and named entity recognition effectively.
- Performs well even on languages it wasn't pre-trained on, showing cross-lingual generalization.

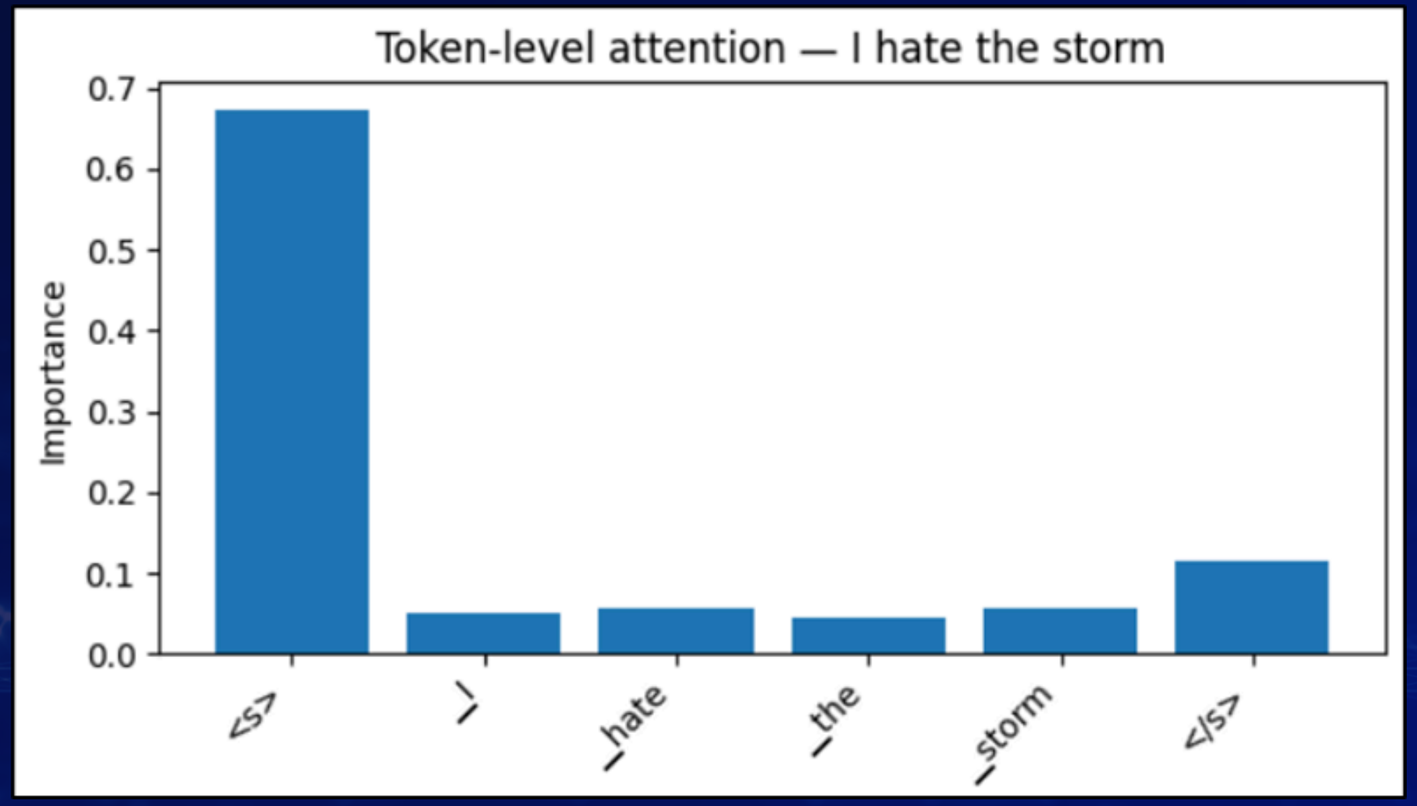
AfroXLMR

- Built on XLM-RoBERTa, trained on 17 African and high-resource languages.
- Uses multilingual adaptive fine-tuning for better zero-shot classification.
- Balances low- and high-resource performance, improving transfer learning across languages.

Insight

- Both models significantly enhance African NLP performance, but AfroXLMR offers broader adaptability across multilingual contexts.

AfroXLMR Token Attention Results

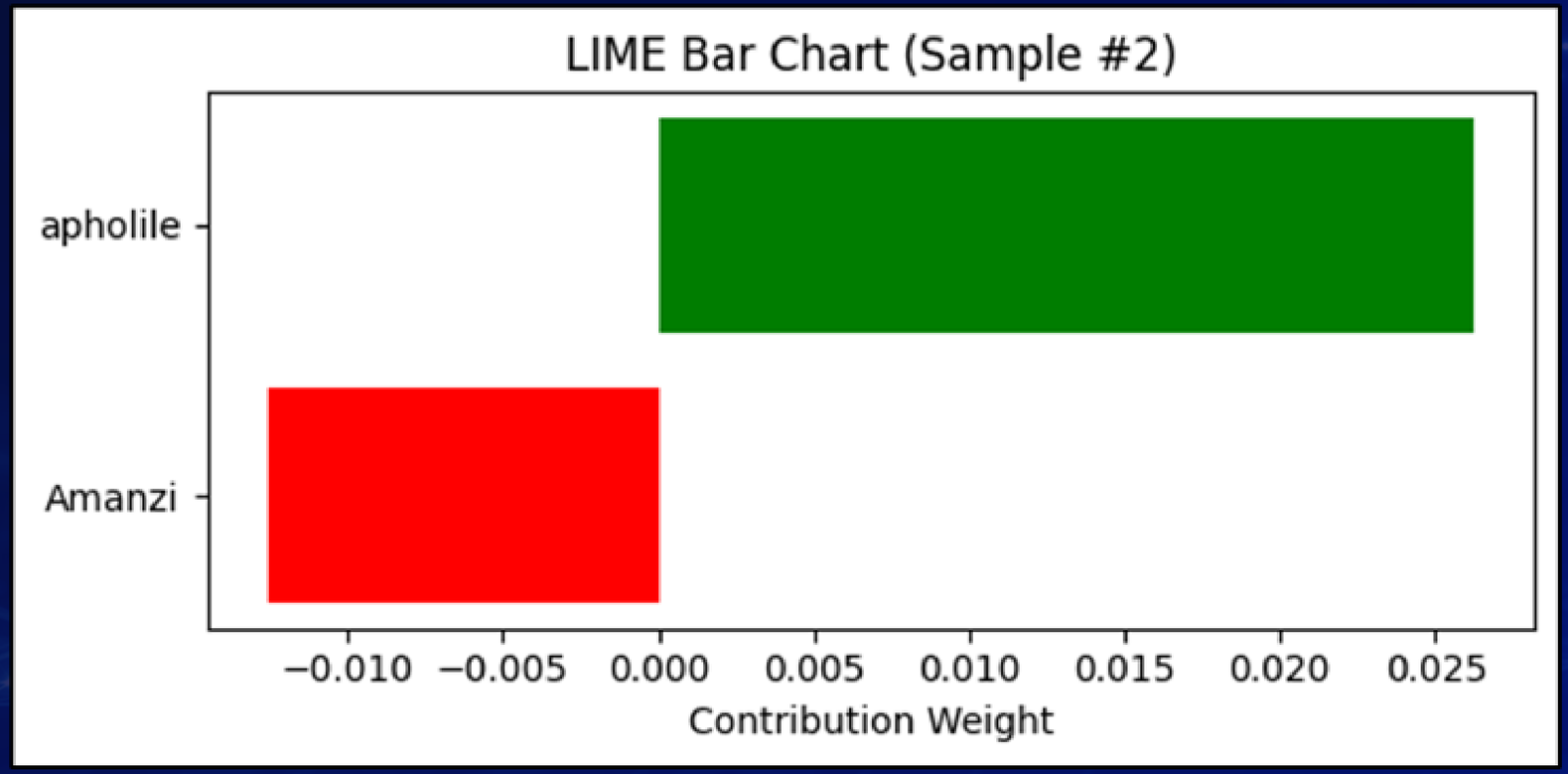


Purpose
To examine AfroXLMR's token-level attention when processing explicit negative sentiment in English sentences, highlighting how the model differentiates sentiment-bearing tokens from neutral ones.
Sentence: "I hate the storm."

Findings

- High attention: The token "hate" received the strongest attention weight, confirming it as the main negative sentiment carrier.
- Low attention: Tokens such as "the" and "storm" were given lower attention, correctly treated as contextual rather than sentiment-driven.
- The visualization indicates that AfroXLMR effectively isolates the emotional core of the sentence, assigning appropriate weights to sentiment intensity.

AfriBERTa Token Attention Results



Purpose

To quantify and visualize how individual tokens contribute to AfroXLMR's sentiment prediction using LIME, providing transparency on which words most strongly influence the model's decision.

Sentence: "Amanzi apholile" → "The water is cold."

Findings

- Positive contribution: "apholile" (+0.026) — strongly supports the predicted sentiment class, indicating positive polarity.
- Negative contribution: "amanzi" (−0.010) — slightly opposes the prediction, showing weaker contextual influence.
- The model's final decision is primarily driven by "apholile", whose higher contribution outweighs the minor negative influence of "amanzi."

AfroXLMR Token Attention Results

Purpose

To assess how AfroXLMR allocates attention to negatively-polarised words, revealing how the model identifies and weighs sentiment indicators across tokens.

Example Context: Negative sentiment expressions such as “don’t”, “all”, and “like”.

Findings

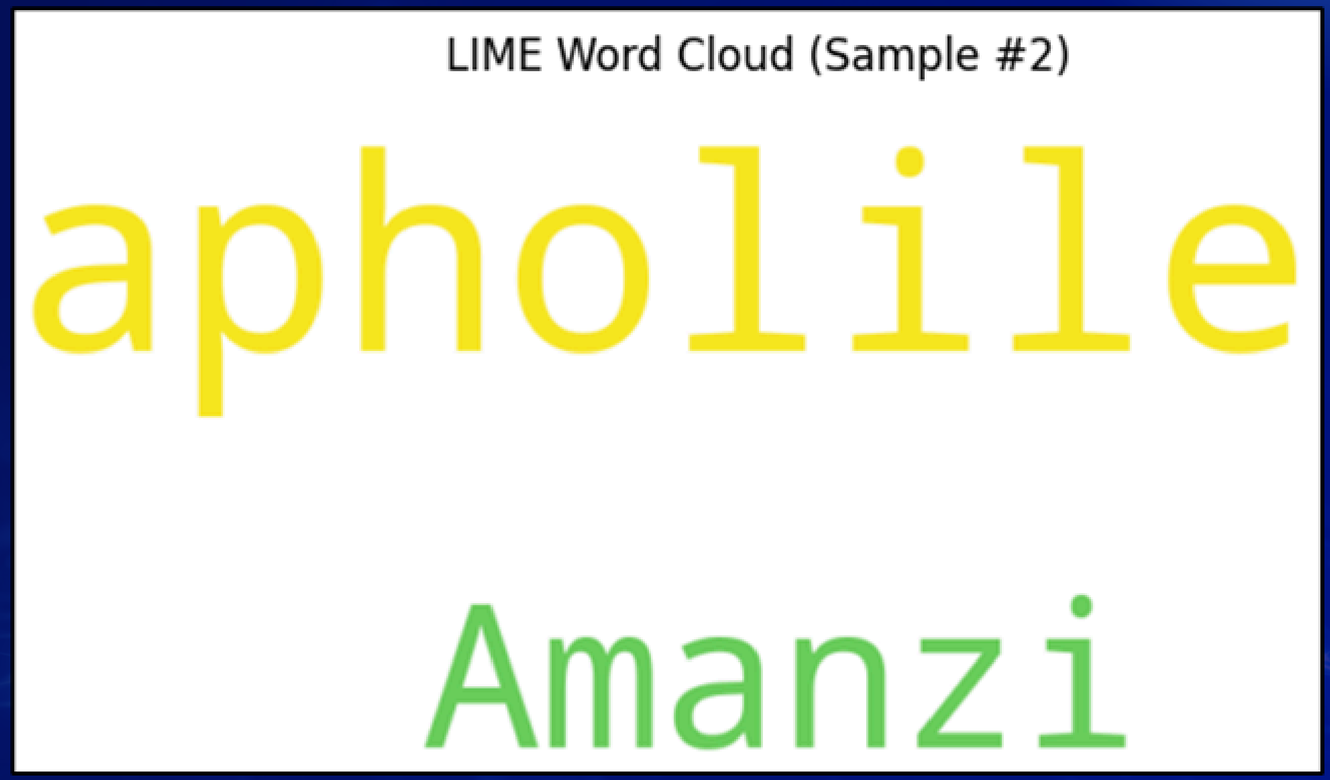
- High attention: “this”, “all”, “don’t”, and “like” — these tokens carry stronger sentiment and received the highest attention weights.
- Low attention: “I” and “at” — minimal polarity contribution but still contextually recognised.
- The variation in word size within the word cloud demonstrates how the model prioritises explicit and contextual negativity indicators.

AfriBERTa Token Attention Results

Purpose
To visualize how AfriBERTa identifies and weights the most influential words in sentiment classification, providing insight into how the model interprets emotional cues in Xhosa sentences.
Sentence: “Amanzi apholile” → “The water is cold.”

Findings

- The token “apholile” appears larger and in yellow, showing it has the strongest contribution to AfriBERTa’s sentiment prediction.
- The token “amanzi”, shown in green, has a weaker influence, contributing less to the model’s overall decision.
- Both tokens contribute positively to the predicted sentiment, with “apholile” identified as the key sentiment indicator.



AfroXLMR Token Attention Results

Purpose

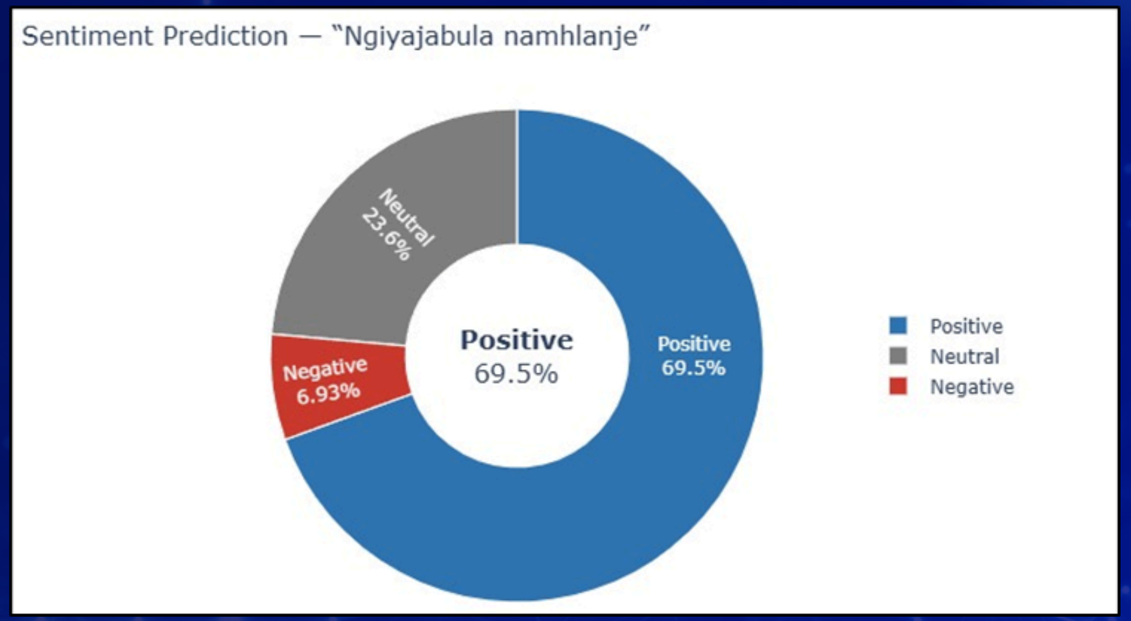
- To evaluate AfroXLMR's confidence distribution in predicting sentiment for a Zulu sentence ("Ngiyajabula namuhla" → "I am happy today").
- To assess how well the model distinguishes emotional polarity in a low-resource, morphologically rich language.

Findings

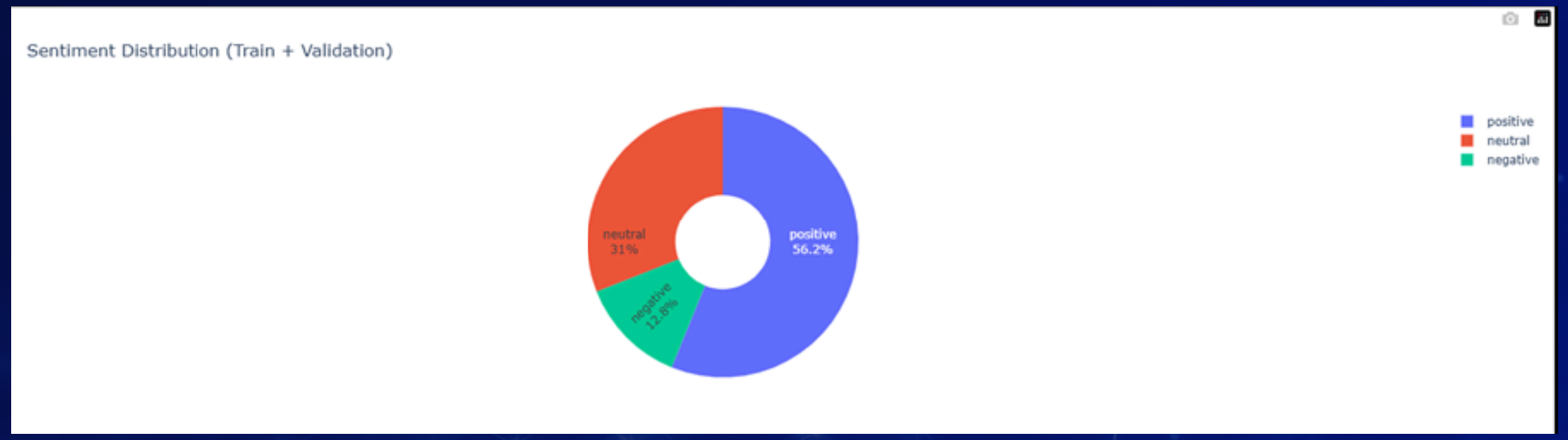
- Positive (69.5%) → Model correctly identifies the happy emotional tone.
- Neutral (23.6%) → Minor contextual uncertainty, likely due to Zulu verb complexity.
- Negative (6.9%) → Very low confusion between opposite emotional cues.

Conclusion

- AfroXLMR effectively captures positive sentiment in Zulu, validating its cross-lingual generalization.
- The small neutral percentage reflects typical uncertainty seen in low-resource language contexts.



AfriBERTa Token Attention Results



Purpose

To illustrate the sentiment distribution across training and validation datasets, highlighting potential class imbalances that may influence model performance, particularly for AfriBERTa.

Findings

- The dataset comprises 56.2% positive, 31% neutral, and 12.8% negative samples.
- The distribution shows a clear imbalance, with positive sentiment being the dominant class.
- Such imbalance can lead to model bias, causing the model to overpredict positive sentiment and underperform on negative examples.

AfroXLMR Token Attention Results

Purpose

- To evaluate how AfroXLMR interprets sentiment in low-resource African languages like Zulu.
- To visually inspect model attention and confirm whether it understands morphological and emotional cues.

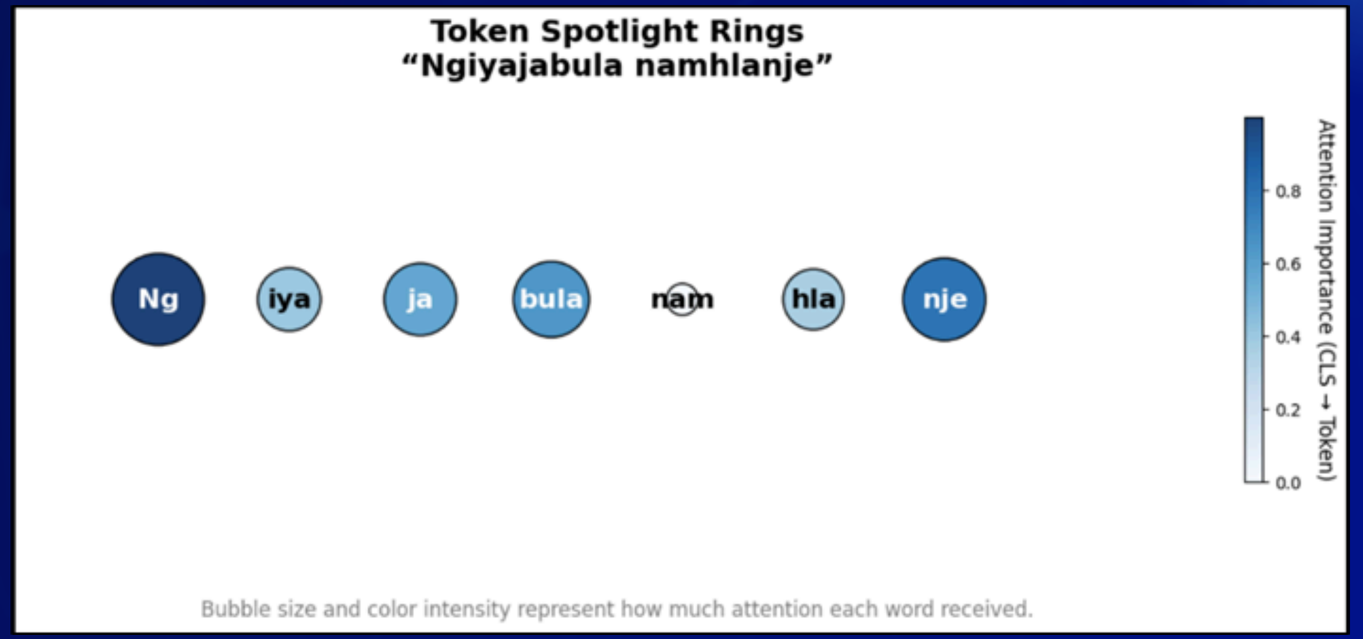
Sentence: “Ngiyajabula namuhla” → “I am happy today.”

Findings

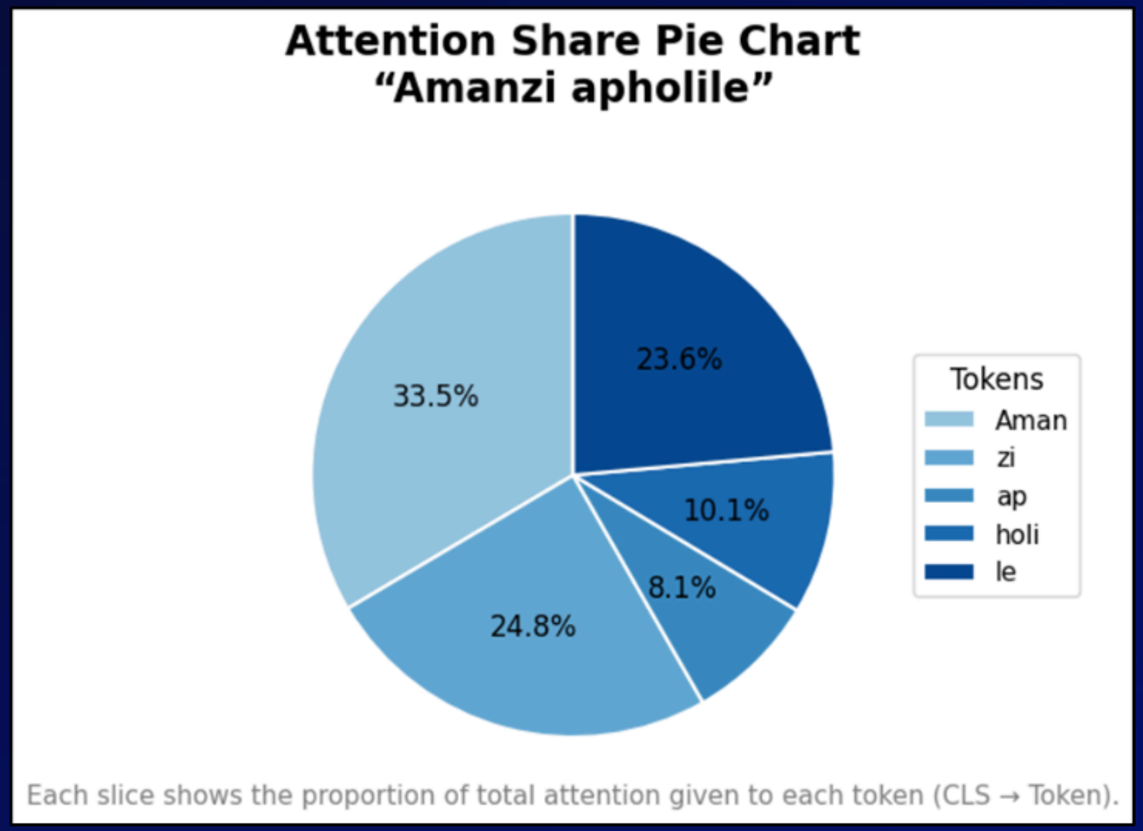
- High attention: “Ng”, “nje” → model focuses on emotion and time indicators.
- Medium attention: “iya”, “ja”, “bula”, “hla” → partial grasp of verb morphology.
- Low attention: “nam” → less focus on locative elements.

Conclusion

- AfroXLMR effectively captures sentiment cues in morphologically rich African languages, demonstrating generalization across complex linguistic structures.



AfroXLMR Token Attention Results



Purpose

To evaluate how AfroXLMR distributes attention across morphemes in a Xhosa sentence to determine whether the model effectively interprets sentiment and semantic cues in low-resource African languages.

Sentence: “Amanzi apholile” → “The water is cold.”

Findings

- The “Aman” token received the highest attention (33.5%), indicating strong focus on the core subject.
- “Zi” and “le” tokens followed with 24.8% and 23.6% respectively, showing balanced model attention across sentence boundaries.
- Moderate attention to “ap” (10.1%) and “holi” (8.1%) suggests that AfroXLMR recognises sub-word morphological cues that convey sentiment (coldness) and polarity.

AfroXLMR BERTViz Attention – Negative Sentiment

Purpose

- To interpret how AfroXLMR processes negative sentiment using BERTViz attention visualization.
- To understand token-level relationships in English and evaluate alignment with human semantic interpretation.

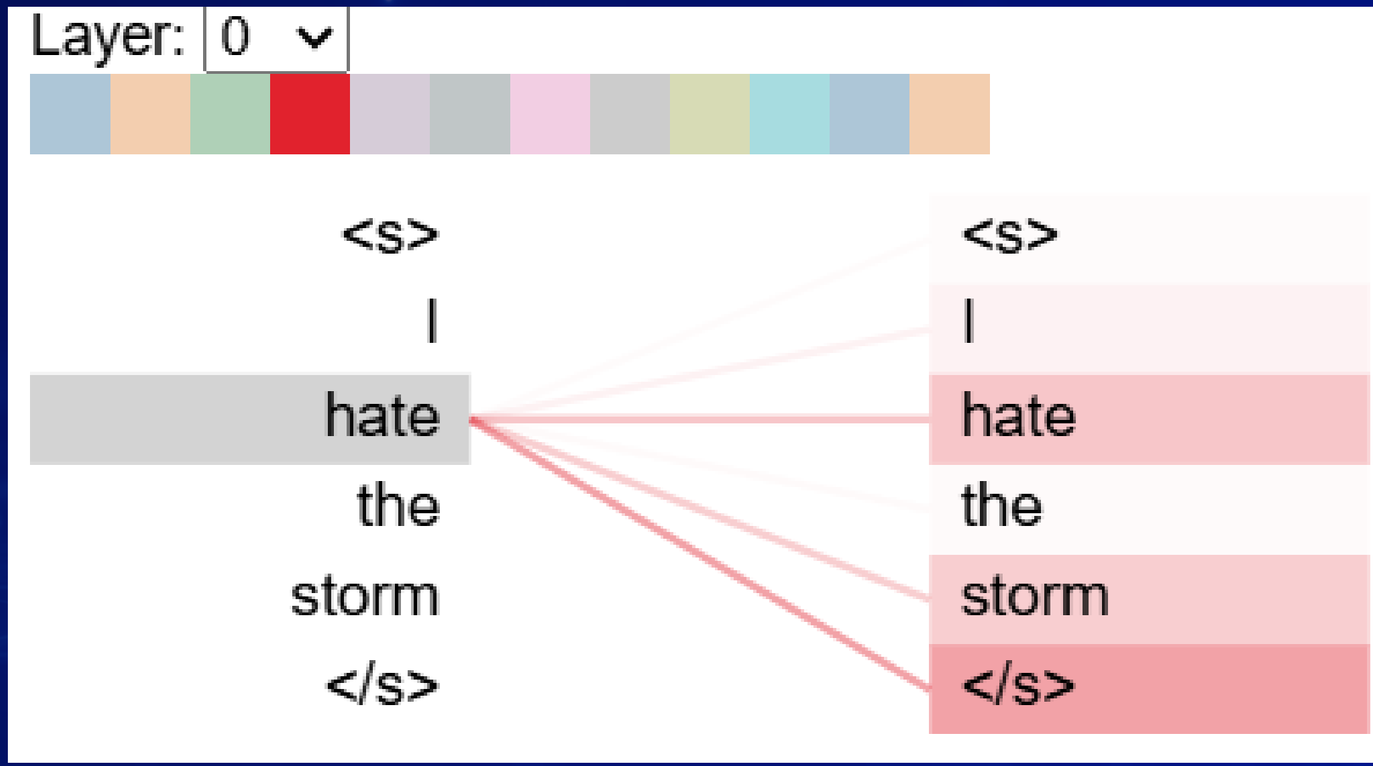
Example Sentence: “I hate the storm.”

Findings

- High attention: “hate” → strongest focus from and to the classification token, showing it drives negative polarity.
- Low attention: “I” and “storm” → support contextual meaning but less influential.
- Symmetrical attention: reveals AfroXLMR understands the relationship between “hate” and “storm” within negative context.

Conclusion

- AfroXLMR effectively isolates sentiment-bearing words and aligns with human interpretation, showing transparent and explainable model behavior.



AfriBERTa Token Alignment Analysis

Purpose

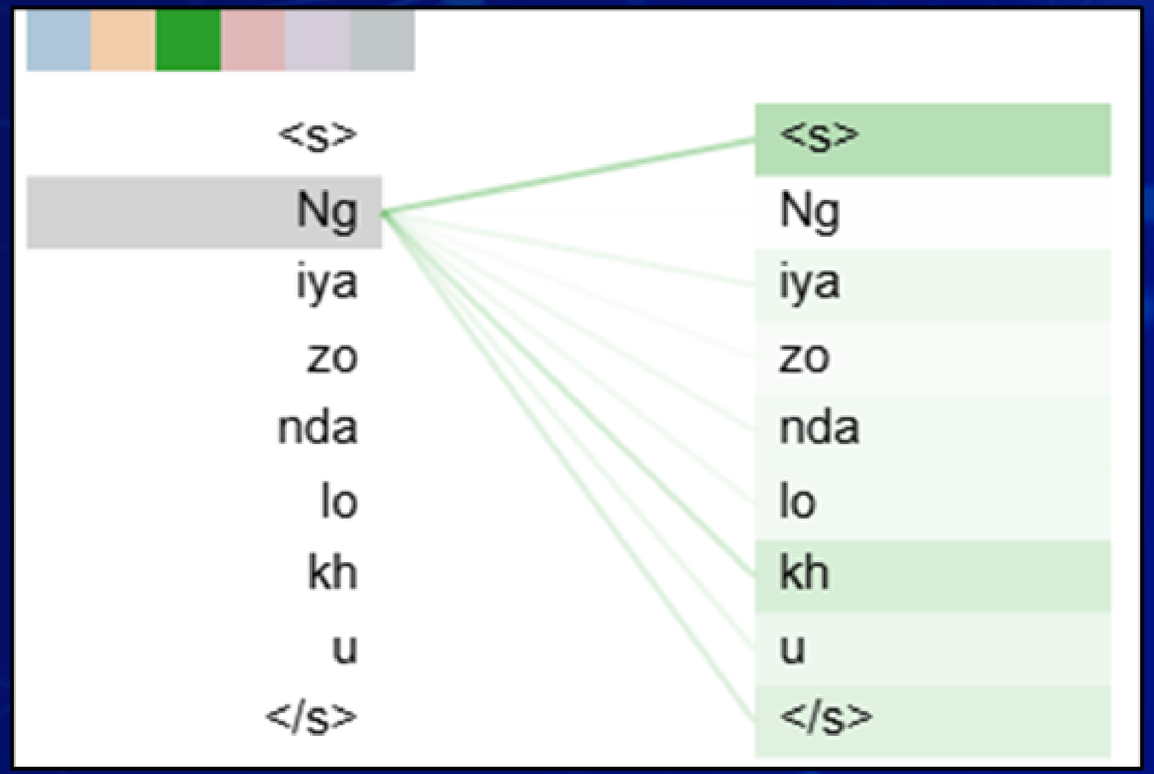
- To examine AfriBERTa's token alignment patterns across Zulu sequences.
- To evaluate how well the model preserves semantic and syntactic consistency during translation or paraphrasing.

Findings

- Cross-sequence alignment:
  - Shows imperfect one-to-one connections between tokens (Ng, iya, xo, nda, lo, kh, u).
  - Indicates AfriBERTa's ability to map semantic and syntactic relationships across slightly varied phrases.
- Self-alignment:
  - Displays perfect one-to-one matching of identical sequences.
  - Reflects stable token preservation and consistent internal structure.

Conclusion

- AfriBERTa demonstrates strong alignment stability and semantic awareness, effectively maintaining token relationships across morphological variations in African languages.



Ensemble Learning Results

- Our group completed ensemble learning using a stacking approach that combined AfroXLMR and AfriBERTa.

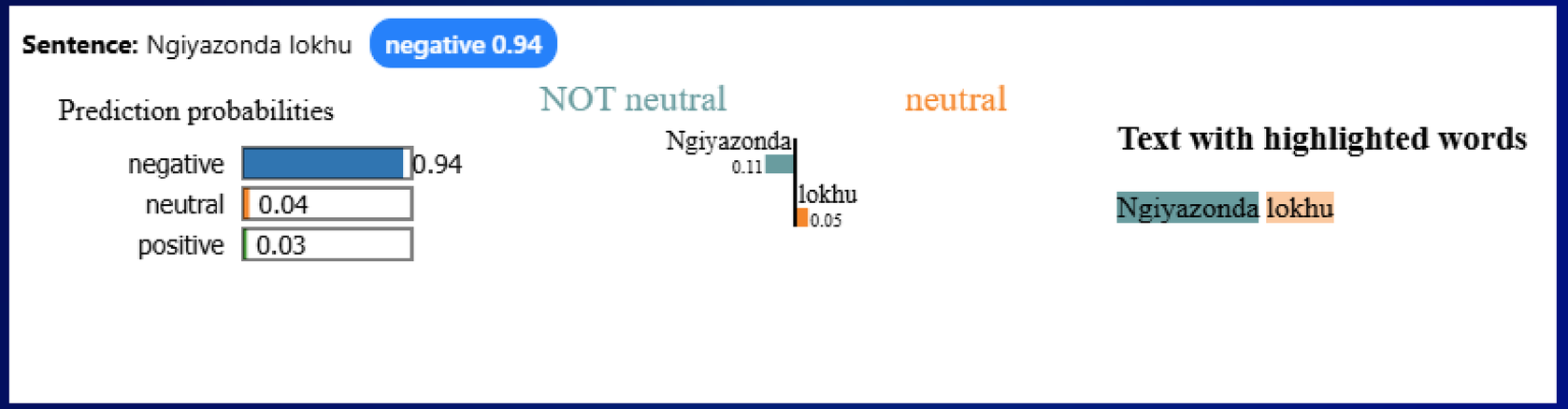
- The meta-learner (Logistic Regression) was trained on the combined probability outputs of both models.
- The ensemble achieved an overall accuracy of 65% and a macro F1-score of 66%.

- These results show that combining the two models improves sentiment classification performance, producing stable and consistent predictions across all sentiment classes.

```
Meta feature shapes: (8681, 6) (1086, 6)

Training meta-learner (LogisticRegression)...

--- Ensemble evaluation on TEST set ---
Accuracy: 0.6464088397790055
F1 macro: 0.6569050056614597

Classification report:

              precision    recall  f1-score   support

           0       0.71      0.81      0.76       183
           1       0.76      0.59      0.67       627
           2       0.47      0.66      0.55       276

    accuracy                           0.65      1086
   macro avg       0.65      0.69      0.66      1086
weighted avg       0.68      0.65      0.65      1086
```

Explainability with LIME

- We used LIME visualisations to understand how the stacked model interprets sentiments at the token level.
- Sentence 1: “Ngiyazonda lokhu” → correctly classified as negative (0.94), with “Ngiyazonda” as the key negative cue.
-



Explainability with LIME

• Sentence 2: “Umlhaba muhle” → correctly classified as positive (0.85), with “muhle” strongly influencing the positive sentiment.

These explainability results confirm that the ensemble can accurately identify sentiment-bearing words in African languages, aligning predictions with the text’s semantic meaning.