Departement Informatika
Department of Informatics

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

University of Pretoria

Dr. Mike Nkongolo Wa Nkongolo

Length (Unlimited). Including the cover page, table of contents, list of tables, list of figures, and reference page(s).

Type: <u>Group Assignment</u> (Not more than 15 students per group). Submit 1 PDF file per group. The page must include all students' full names, student numbers, and the group Nickname.

Submission Date: 25 October 2025

Submission Method: Online ClickUP + Group Presentation

Title: Expanding a Multilingual Lexicon for Low-Resource Language Processing

Problem Statement

Although multilingual sentiment analysis systems have achieved strong results for widely spoken languages, many African languages with limited resources remain underrepresented. This restricts access to essential language technologies for large populations, particularly in their native languages. Multilingual pre-trained language models (PLMs) provide an opportunity to extend sentiment analysis to low-resource settings, but South African low-resource languages still lack sufficient resources for full support.

This assignment addresses this gap by expanding an existing multilingual sentiment lexicon composed of French and Ciluba to include South Africa's major indigenous languages, such as Sesotho, Sepedi, Zulu, Xhosa, Shona, and English or Afrikaans. Using existing corpus (e.g., ShonaSenti), sentiment-bearing words will be extracted through Pointwise Mutual Information (PMI) and aligned with the multilingual lexicon. Retrieval-Augmented Generation (RAG), together with Large Language Models (LLMs) such as Google Translate and ChatGPT, will be applied to refine ambiguous entries and support efficient annotation.

AfroXLMR and AfriBERTa will be evaluated as the main PLMs for sentiment classification, with their ensemble learning explored to determine whether combining PLM models improves performance. Explainable AI (XAI) techniques will further be used to interpret model predictions and provide insights into how sentiment is determined. The overall goal is to improve sentiment analysis resources for low-resource African languages and demonstrate how multilingual PLMs, hybrid annotation, and explainability methods can support this advancement.

Given Lexicon

The base lexicon includes the following columns:

1. **Ciluba** – a low-resource language spoken in the Democratic Republic of Congo (DRC).
2. **French** – translations of the Ciluba entries.
3. **Score** – a sentiment score ranging from -9 (strongly negative), 0 (neutral), to 9 (strongly positive).
4. **Sentiment** – sentiment class (negative, positive, neutral).
5. **Nature** – part-of-speech (POS) tags (e.g., verb, noun, number).

Tasks

A. Lexicon Expansion

Students must expand the lexicon by adding additional columns for low-resource South African languages (e.g., Sepedi, Zulu, Xhosa, Afrikaans, Shona, and more).

- Expansion will be done by first translating French words into English using LLMs, and then translating English words into the target low-resource languages.

- The expanded lexicon must include both the original and additional columns (no original columns should be deleted).

- Students must also report the expansion algorithm or framework used.

B. Corpus-Based Enrichment

- Use an existing labeled corpus (e.g., ShonaSenti or more).

- Apply text cleaning and tokenization, then use PMI to identify candidate sentiment-bearing words.

- Translate identified words into English with Google Translate API. Perform the same process on the lexicon by converting English words into low-resource South African languages.

- Compare sentiment polarity assignments across corpus and lexicon entries.

    - If polarity matches → retain the sentiment.

    - If polarity differs → review ambiguous words using RAG and LLMs (e.g., ChatGPT prompts requesting POS tags and polarity values).

C. Model Training with PLMs

- Use **AfroXLMR** (a multilingual model based on XLM-RoBERTa) and **AfriBERTa** (trained on African languages).

- Since these models can behave like "black boxes," employ Explainable AI (XAI) techniques to analyze predictions and identify the features influencing sentiment classification.

- Explore **ensemble learning** to assess whether combining AfroXLMR and AfriBERTa improves classification accuracy.

D. Research Questions

This assignment is guided by the following research questions:

1. How can a multilingual sentiment lexicon be expanded to include low-resource languages using corpus-based extraction, cross-lingual mapping, and Retrieval-Augmented Generation (RAG)?

2. What is the effect of integrating low-resource languages into the lexicon on the coverage and accuracy of sentiment analysis?

3. How effective are AfroXLMR and AfriBERTa for aspect-based sentiment analysis with the expanded lexicon?

4. How does ensemble learning of AfroXLMR and AfriBERTa influence sentiment classification performance?

Approach

- **Extraction**: Identify sentiment-bearing words using PMI.

- **Mapping**: Use Google Translate API for cross-lingual alignment.

- **Validation**: Compare sentiment assignments across corpus and lexicon entries.

- **Disambiguation**: Use RAG to retrieve contextual examples and LLMs for assigning POS and polarity values.

- **Classification**: Perform aspect-based sentiment analysis using AfroXLMR and AfriBERTa.

E. Text Normalization and Tokenization

To ensure consistency and reliability, normalization will be applied:

1. **Lowercasing** – convert all text to lowercase.

2. **Spelling correction** – hybrid approach using ChatGPT, Google Translate API, and manual review by native speakers.

3. **Punctuation removal** – remove commas, periods, exclamation marks, etc.

4. **Whitespace and special character cleaning** – remove extra spaces, tabs, and non-alphanumeric characters.

Tokenization:

- **Whitespace-based tokenization** for lexicon building.

- **Subword tokenization (SentencePiece/BPE)** for transformer-based models (AfroXLMR and AfriBERTa).

F. Statistical Extraction

- Apply frequency analysis and PMI to measure word-sentiment association.

- Identify words that co-occur with sentiment indicators in the corpus to strengthen lexicon entries.

G. Explainable AI (XAI) and Model Interpretation

Given the complexity of AfroXLMR and AfriBERTa:

- Use **attention visualization** to highlight influential words or phrases in sentiment predictions.

- Inspect attention weights to check whether the models focus on relevant sentiment-bearing words or irrelevant context.

- Provide transparent explanations to improve trust in model decisions.

Rubric

Category	**Description**	**Weight (%)**
Presentation & Clarity	Report structure, writing quality, formatting, flow, and clarity of explanations. Figures/tables well-labeled and easy to follow.	5%
Problem Understanding & Motivation	Clear explanation of the problem, context of low-resource languages, and relevance of multilingual sentiment analysis.	10%

Lexicon Expansion (Task A & B)	Correct implementation of lexicon expansion: translation pipeline (French → English → South African languages), incorporation of corpus data, and alignment of sentiment labels. Includes description of algorithm/framework used.	20%
Corpus Processing & PMI Extraction	Application of text normalization, tokenization, and Pointwise Mutual Information (PMI) to extract sentiment-bearing words. Demonstrates understanding of statistical methods.	15%
Use of LLMs & RAG	Effective use of Google Translate, ChatGPT, and Retrieval-Augmented Generation (RAG) for disambiguation and refinement of lexicon entries.	10%
Model Implementation (AfroXLMR & AfriBERTa)	Correct application of AfroXLMR and AfriBERTa for sentiment classification, with clear description of model pipeline.	10%
Ensemble Learning & Performance Analysis	Evaluation of ensemble methods (AfroXLMR + AfriBERTa), reporting performance metrics, and analyzing improvements.	10%
Explainable AI (XAI) Application	Application of XAI methods (e.g., attention visualization, feature importance) to interpret predictions, showing transparency of model decisions.	10%
Group Presentation & Reflection	Discussion of strengths, weaknesses, limitations, and possible improvements of the approach. Relates findings back to low-resource language challenges.	10%
Total		**100%**