UNIVERSITY OF THE WESTERN CAPE

Private Bag X17, Bellville 7535, South Africa
Tel: +27 21 959 2911 | www.uwc.ac.za

Annual Progress Report - 2025

Department of Computer Science

Student Name:	Amahle Dlamini
Student Number:	3856712
Supervisor:	Dr. James Okafor
Department:	Computer Science
Programme:	MSc Data Science
Reporting Period:	1 January 2025 - 31 December 2025
Date Submitted:	20 January 2026

1. EXECUTIVE SUMMARY

This report summarises research progress for the 2025 academic year. The study investigates the application of Natural Language Processing (NLP) techniques for multilingual sentiment analysis of South African social media data, with a focus on isiZulu, isiXhosa, and Afrikaans text. Key achievements include the development of a novel tokenisation approach for agglutinative languages, collection of a 50,000-tweet annotated dataset, and a baseline BERT model achieving 78.4% accuracy on the multilingual sentiment classification task.

2. COMPLETED ACTIVITIES

- [x] Completed comprehensive literature review on NLP for low-resource African languages
- [x] Developed custom tokeniser for isiZulu and isiXhosa (handles noun class prefixes and agglutination)
- [x] Collected and annotated 50,000 tweets in three languages using crowd-sourced annotators
- [x] Trained baseline multilingual BERT model (mBERT) achieving 78.4% accuracy
- [x] Presented poster at the AfricaNLP Workshop (co-located with ICLR 2025)
- [x] Completed Research Ethics online training module (CITI Program)
- [x] Submitted abstract to the African Conference on Computational Linguistics (AfriCCL 2026)

3. CHALLENGES

The primary challenge has been sourcing high-quality annotators for isiZulu sentiment. Many social media posts contain code-switching between English and isiZulu, which complicates annotation consistency. Inter-annotator agreement (Fleiss kappa) was initially 0.62 but improved to 0.78 after two rounds of annotator training. Additionally, Twitter API access changes in mid-2025 required migrating data collection to alternative APIs.

4. PLAN FOR NEXT PERIOD

- > Fine-tune AfroXLMR (African-specific transformer) and compare with mBERT baseline
- > Implement custom attention mechanism for code-switched text
- > Expand dataset to 100,000 annotated examples
- > Begin thesis writing - Chapters 1-3 (Literature Review, Methodology, Data Collection)
- > Submit journal paper to Natural Language Engineering