Contribution ID: **11** Type: **Talk**

A Retrieval-Based NLP Chatbot for Setswana Proverbs and Idioms Interpretation Using TF-IDF and Cosine Similarity

Indigenous proverbs and idioms preserve cultural knowledge, moral values, social wisdom and linguistic identity, yet many African languages remain underrepresented in natural language processing resources and applications. This study presents the development of a retrieval-based cultural chatbot for interpreting Setswana proverbs and idioms, leveraging a bilingual Setswana–English knowledge base. The proposed system applies text preprocessing, Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation and cosine similarity to match user queries with the most relevant proverb or idiom in the dataset. Unlike data-intensive deep learning approaches, the model is suitable for low-resource language contexts because it does not require a large annotated corpus. The chatbot was evaluated using Top-1 Accuracy, Top-3 Accuracy, Mean Reciprocal Rank and Mean Similarity Score to assess its retrieval effectiveness and ranking quality. The results indicate that the system performs strongly in retrieving relevant cultural expressions, with high Top-1 and Top-3 accuracy scores, demonstrating its practical value for language learning, cultural interpretation and indigenous knowledge preservation. Visualisations such as performance metric charts, similarity score distributions and PCA representation were used to analyse the model's behaviour and the structure of the proverb representations. The study contributes to African language technology by demonstrating how lightweight NLP techniques can support digital access to Setswana cultural expressions. Future work may expand the dataset, incorporate cultural explanations and thematic annotations, and explore multilingual semantic embeddings or transformer-based models for deeper meaning-based retrieval.

Primary author: Dr MASETHE, Hlaudi (Tshwane University of Technology)

Presenter: Dr MASETHE, Hlaudi (Tshwane University of Technology)

Track Classification: DIRISA