SUBWORD-AWARE NEURAL CLASSIFICATION OF SENTIMENT FOR THE SESOTHO LANGUAGE

Innocentia Ledimo **Khotso Bore** **Nokubonga Cele**
Department of Computer Science, University of Pretoria
COS760 – Natural Language Processing
{u21556564, u19180642, u25717342}@tuks.co.za

Abstract

Neural language models have become central to many NLP tasks like translation and sentiment analysis. To deal with rare or unknown words, these models often break words into smaller parts — a process known as subword tokenization. While this helps in many cases, it can cause problems when applied to African languages like Sesotho and Setswana, where word structure is more complex. In these languages, subword tokenization may lose important meaning, making it harder to understand what's really being said. Our project explores how this affects sentiment classification. We build a neural model that works with subword tokens and uses STF-IDF to classify text at the word level. We then compare its performance to traditional models like linear regression, as well as pre-trained multilingual models. By doing this, we aim to find out which approaches work best for handling sentiment in low-resource African languages — and how to reduce the loss of meaning along the way.

1 Introduction

In recent years, neural language models have become the cornerstone of Natural Language Processing (NLP), enabling advances in tasks such as machine translation, sentiment analysis, and information retrieval (**?**)(**?**) Central to many of these models is the technique of subword-tokenization, which breaks down words into smaller units to handle rare or unseen terms. However, this technique presents challenges when applied to African languages(**?**). Subword-tokenization can distort meaning and lead to semantic loss, hindering the performance of NLP systems for these languages.

This research addresses the limitations of subword tokenization in neural language models, particularly in the context of sentiment classification for low-resource African languages like Sesotho and Setswana. These languages are underrepresented in both annotated datasets and in the design of existing pre-trained models.

Our objective is to develop a neural model that utilizes subword tokenization for input but performs classification at the word level. By integrating subword-aware embeddings with traditional techniques such as TF-IDF, and comparing our model's performance with pre-trained multilingual models, we aim to demonstrate more effective sentiment classification for African languages.

Our work directly addresses the issue of under representation in NLP. African languages are vastly under served in mainstream language technologies, leading to a digital divide that marginalizes millions of speakers. By focusing on low-resource languages and developing models that account for their unique linguistic features, we aim to improve fairness and accessibility in NLP.

2 Background

Natural Language Processing (NLP) for African languages is gaining momentum, yet remains under-resourced compared to high-resource languages. Recent work has introduced new datasets and modeling techniques, but significant limitations still occur. For instance, the SAfriSenti corpus was developed to support sentiment analysis across several South African languages (e.g. Setswana, Sesotho, Sepedi, isiXhosa, and isiZulu)(**?**), highlighting the effectiveness of multilingual pre-trained language models such as AfroXLMR and AfriBERTa when fine-tuned on related languages (**?**). Tokenization strategies have also been evaluated in the context of machine translation. Rajab (2022) compared subword BPE approaches and demonstrated that tokenization quality directly impacts model performance on African languages such as Setswana and isiZulu. SentencePiece tokenization showed improved BLEU scores due to its language-agnostic encoding, which better accommodates the agglutinative structure common in African languages (**?**). Despite these advancements, major gaps remain. Even well-known models like mBERT and XLM-R have limited coverage and training data for African languages(**?**). Moreover, many datasets, such as SAfriSenti, while useful, rely heavily on social media content and may not capture the full linguistic richness of African languages. Prior research in African NLP has quite a huge difference, but a large gap still exists. Tokenization algorithms, pretrained models, and sentiment datasets often assume linguistic features aligned with European languages. This results in reduced effectiveness for languages such as Setswana and Sesotho. Existing datasets are limited in size and linguistic variation, often made from narrow domains like Twitter, which can introduce bias and restrict generalizability (**?**).

3 Methodology

As previously mentioned, the aim of our study is exploring how effective subword-aware neural language models are, specifically for AfroXLMR(**?**) in sentiment analysis with deep focus on African languages that are considered low-resource.With another goal of providing meaningful evaluation, we compared the performance of AfroXLMR with traditional classification models of which include TF-IDF with logistic regression, and a custom subword TF-IDF(STF-IDF) model.

Our methodology is structured in the following manner: Data selection, Preprocessing, Feature engineering, model training, and Evaluation.

To elaborate further on the processes:

1. Dataset Collection and Selection

   We used the two prescribed datasets as our main sources of data:

   - Sesotho subset of the SAfriSenti Corpus. It contained sentiment labeled tweets.[1]
   - Sesotho news headlines dataset, locally compiled. This dataset was stored in a custom format, unlabeled consisting of sentences as well as the sentiment integers.

   Choosing the SAfriSenti dataset was influenced by the fact that it was curated just for sentiment analysis as depicted in its format as well as the contents. It provides content that is user generated in real world. It is also inherently noisy, informal and short, which are characteristics that tend to pose a challenge on standard nlp techniques. However, these are also good characteristics providing an ideal test environment for subword based models with a focus on handling out-of-vocabulary(OOV) terms and spelling variations. The Sesotho news headlines dataset provides a more formal domain complementing the informal tweets dataset, which allows us to evaluate domain robustness of our model.

   We have selected the Sesotho language because it has proven to be less covered in the mainstream multi lingual sentiment dataset, thus making it an ideal case for cross lingual transfer learning.

2. Preprocessing

   Preprocessing a dataset is a critical step for both neural and classical approaches to this study. For the traditional aspect, we used a rule-based cleaning approach. This includes the following operations:

   - User Mentions: The SAfrisenti tweet had "@user" texts which made sense considering they are tweets but they were unnecessary as per the work and goals we wanted to achieve, thus they had to be removed, as well as any existing URLs as they carry little semantic value and aid with noise.
   - Hashtags: punctuations and numbers were eliminated promoting focus on the linguistic content of the sentence.
   - Lowercasing: For standardization.
   - Emoji and emoticons: Emojis are not consistently tokenized across models, thus done through stripping unicodes.
   - Whitespace normalization: This ensures that the content or text is split in a consistent manner, consistent token. This is especially important for tokenizers such as TF-IDF as they are whitespace-sensitive.
   - Label Standardization: There were also label mismatches between datasets which needed to be corrected before combining them. The SAfrisenti-Corpus dataset had sentiment labels as text. “positive,” “negative,” “neutral” and the Sesotho News headlines had integer la-

[1] However, it has recently been removed from Github.

bels -1,0,1 for negative, neutral, and positive sentiment respectively. We standardized these to fit as integer classes 0,1,2 for negative, positive, and neutral sentiment respectively
Many African languages use diacritics and affixes and are very morphologically rich, thus carry semantic meaning. By intentionally retaining these diacritic characters as language specific tokens, we preserve the model's ability to learn meaningful subword patterns and maintain the morphology.

3. Feature Engineering

   Fine Tuning AfroXLMR: Subword-Aware Transformer Model

   Our main model, AfroXLMR was pretrained on 17 African languages. It uses subword tokenization through sentencepiece subword splitting . Thus a good model to get started with.

   We fine-tuned our model for sequence classification task using the Hugging Face Trainer API (WandB):

   - The auto-tokenizer split the text or input into subword units and then mapped them to token IDs.
   - We combined the Sesotho news headlines and the SAfriSenti Sesotho dataset and fine-tuned the model for 3-5 epochs.
   - A classification head with softmax output computes probabilities for sentiment classes.

   AfroXLMR has been optimized specifically for African languages. Its vocab is adapted to low-resource African subwords, which makes it extremely effective on code-switched texts and morphologically rich texts. It has the ability to generalize unseen but linguistically similar languages such as Sesotho which just makes it ideal for our study.

4. Baseline and Comparative models

   TF-IDF and STF-IDF logistic regression models were used as baseline models to set a benchmark for comparison for our neural models.

   We Trained two types of neural models on our combined dataset on classification. We identify these models as the Test models and Competitive models.

   For the Test models, the architecture is as follows.

   - Input: The input layer would accept subword embeddings or STF-IDF vectors
   - Hidden layers: 2 hidden layers with dimensions of 128.
   - Output: An output layer with 3 dimensions for outputs of the 3 classes. Positive, Negative and Neutral
   - Optimization: Adam optimization
   - Loss function: CrossEntropyLoss

   The Test models are used as standard indications of potential performance of STF-IDF. From these we developed models that would accept subword embeddings tokenized using BPE to act as a baseline, comparing the effectiveness of subword tokenization to STF-IDF. The models accepting STF-IDF vectors as inputs used different tokenization methods (N-gram, BPE, Word-Piece)

   Lastly the Competitive models. Somewhat larger, these models were made to compete against the AfroXLMR-Large model. The hope was that larger models would potentially have better performance.

   The model details being:

   - Input: The input layer would accept STF-IDF vectors
   - Hidden layers: 7 hidden layers with varying dimensions during examination. Hoping more layers would lead to better performance.
   - Output: An output layer with 3 dimensions for outputs of the 3 classes. Positive, Negative and Neutral
   - Optimization: Adam optimization
   - Loss function: CrossEntropyLoss

4 Experiments And Results

1. Evaluation Strategy

   To evaluate our model performance, we used the following metrics.

   - Accuracy: This metric depicts the overall correctness of predictions.

- Precision: Measures the accuracy of positive predictions. It aims to maximize true positives and minimize false positives.
- Recall: Identifies overall positive predictions. It aims to maximize true positives and minimize false negatives
- F1-Score: The F1-score is the harmonic mean of both and balances both precision and recall. F1 gives a better sense of real performance especially on imbalanced datasets like ours.

2. Transformer Performance Evaluation

The AfroXLMR model achieved a validation accuracy of 60% on the Sesotho dataset. While it performed well(all things considered) on the dominant negative class, the performance on the minority classes -neutral and positive is lower. The F1-Score indicates the model's moderate aggregate reliability, with precision being the lowest metric.

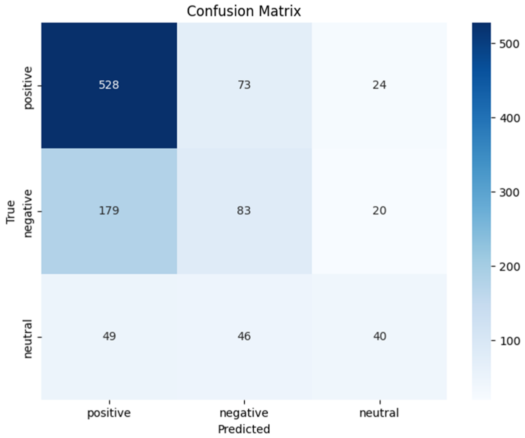
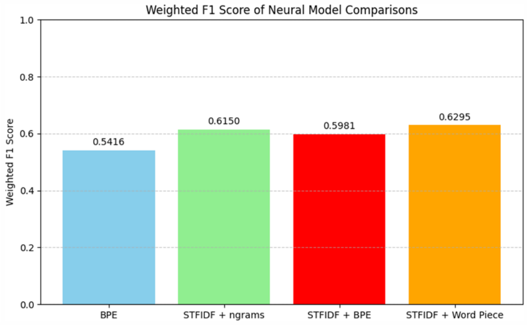


Figure 1: AfroXLMR Confusion Matrix

Adding on to our evaluations, the model's confusion matrix reveals that the model does demonstrate a strong performance in identifying positive sentiments, correctly classifying 528 examples. However, it under performs significantly on negative and neutral classes, with many negative samples incorrectly labeled as positive. Similarly, the neutral class is also incorrectly classified, likely due to its under-representation in the dataset. This imbalance highlights a bias toward the positive class, which skews the decision boundaries of the model.

3. Test Model Analysis

Results from the Test models show weighted-average F1-scores of 59% and greater for the STF-IDF models. The best being the STF-IDF Word-Piece model. These scores are considerably higher than the test model which uses standard embeddings tokenization with BPE. This model scored a weighted-average F1-score of 54%



4. Competitive model Hyperparameter tuning

For the Competitive models we extensively tuned the hyper-parameters like training epochs, learning rate and hidden layer dimensions to find the best possible model for each subword tokenization method (N-gram, BPE, Word-Piece) combined with TF-IDF

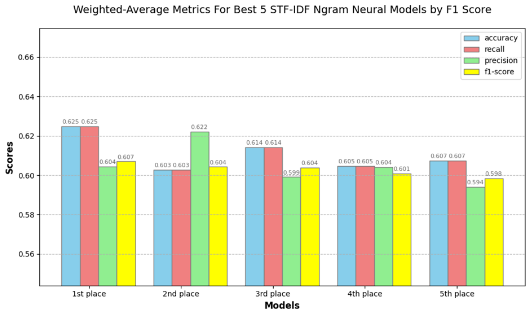


Figure 2: The best N-gram model configuration had dimensions:256, Epochs:20 and Learning rate: 0.001

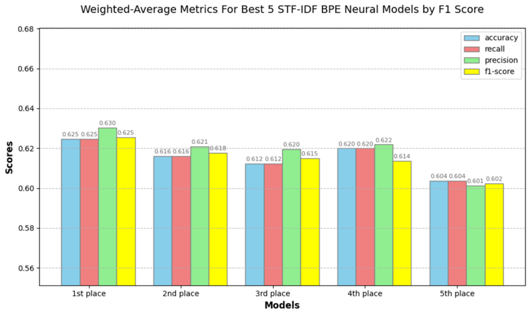


Figure 3: The best BPE model configuration had dimensions:128, Epochs:10 and Learning rate: 0.001

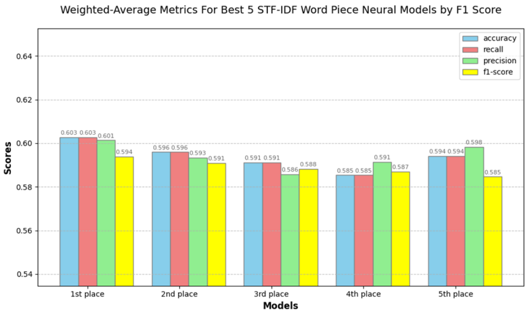


Figure 4: The best Word-piece model configuration had dimensions:256, Epochs:15 and Learning rate: 0.001

5. Comparative model analysis

   From our tuned Competitive models on 3 different subword tokenization methods (N-gram, BPE and Word-Piece) combined with TF-IDF. The STF-IDF neural models scored F1-scores above 59%, the best model being the STF-IDF BPE model scoring 62.55%. All considerably higher than any of the comparative models which scored 46.67%, 45.14% and 46.77% for the TF-IDF + Logistic regression, BPE STF-IDF + Logistic regression and AfroXLMR-large model respectively

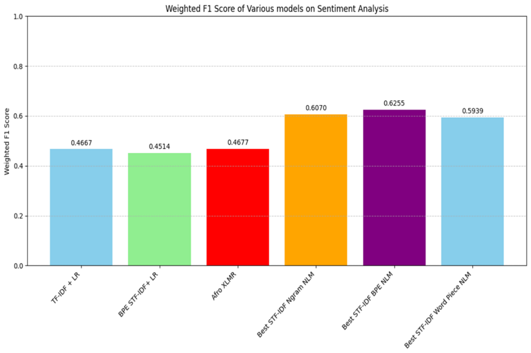


Figure 5: Model comparisons

5 Reflections and Discussion

What worked.

Model selection: The choice of AfroXLMR was appropriate for multilingual tasks and low-resource languages, ensuring coverage of African languages that are often underrepresented in mainstream NLP models.

Data sourcing: The effort to increase the dataset size from 3000 to 5000 improved model performance potential, showing adaptability and resourcefulness.

Platform utilization: Google Colab provided accessible GPU resources for model training without additional cost.

What Didn't Work.

Data quantity constraints: The limited number of samples was still small for transformer models, which typically require significantly larger datasets to generalize well. This affected the model's overall accuracy and robustness.

Computational Limits: Training on Google Colab faced repeated interruptions, session limits, and runtime errors. These factors restricted experimentation, hyperparameter tuning, and thorough model validation.

Lessons Learned: Need for scalable infrastructure: For large-scale models like AfroXLMR, more stable and scalable compute resources

Importance of data quality and quantity: Sufficient and clean data significantly impacts model performance. Future projects should allocate more time for data collection, cleaning, and augmentation.

Experimentation limitations: Resource-constrained environments limit the ability to perform necessary iterations, affecting the ability to explore different model architectures or training strategies. Possible Extensions or Improvements

6 Conclusion

This research contributes both the methodological and practical insights and solutions to the African languages in NLP. Our research findings and experiments demonstrate that innovative hybrid approaches can effectively bridge the gap between traditional methods and modern neural architectures, providing a promising path for African language technology development across the African continent. The depicted success of our STF-IDF neural models, which achieved F1-scores of 62.55% compared to 46.77% for AfroXLMR large, challenges the notion that larger pre-trained models are always superior for low-resource languages. By developing effective NLP tools for Sesotho, this work takes an effective step toward reducing the digital divide that has marginalized millions of African language speakers. While challenges such as data scarcity and computational constraints remain, our research establishes that resource-efficient hybrid approaches may be strong candidates and more viable than purely transformer-based solutions for organizations working with limited resources. As we move forward, the success of our approach promotes NLP investigation across other African languages and highlights the need for collaborative approaches to develop larger, more varied datasets for continuous advancement in African NLP research.

References

S. J. Mielke, Z. Alyafeai, E. Salesky, *et al.*, “Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp,” *arXiv preprint arXiv:2112.10508*, 2021.

S. Mesham, L. Hayward, J. Shapiro, and J. Buys, “Low-Resource Language Modelling of South African Languages,” Apr. 2021. arXiv:2104.00772 [cs].

S. M. L. H. J. S. J. Buys, “Low-resource language modelling of south african language,” *arXiv preprint arXiv:2104.00772*, vol. 1, p. 10, 2021.

R. Mabokela, “Safrisenti-corpus: A multilingual sentiment corpus for south african under-resourced languages,” 2025. Repository for sentiment analysis in South African languages.

K. R. Mabokela and M. P. T. C., “Advancing sentiment analysis for low-resourced african languages using pre-trained language models,” *PLOS ONE*, p. 37, 2024.

J. Rajab, “Effect of tokenisation strategies for low-resourced southern african languages,” in *AfricaNLP Workshop at ICLR 2022*, p. 8, 2022.

Davlan, “Afroxlmr-large: A multilingual language model for african languages,” 2022. Repository for AfroXLMR-large, adapted for African languages.

X. Huang, L. Xing, F. Dernoncourt, and M. J. Paul, “Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition,” *arXiv preprint arXiv:2002.10361*, Mar. 2020.