UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

COS781 Project Proposal

October 2024

Clustering Analysis on Sentiment Tagged Parallel Corpus for Luganda and Swahili

By

Arno Jooste - u21457451

Supervisors:

Prof. Vukosi Marivate

Ms. Seani Rananga

Dr. Abiodun Modupe

Research Questions

- **What is the problem being solved?**
  Recent years have indicated a lack of work on African languages in the field of sentiment analysis [1]. Therefore, I aim to investigate the use of clustering algorithms in low-resource languages, particularly Luganda and Swahili. The challenge might be to successfully apply unsupervised learning algorithms, such as clustering, to sentiment tagged corpora in languages with limited language resources.

- **Why is it interesting?**
  Clustering enables us to identify natural groupings within the data without relying on predetermined classes. Investigating this approach for low-resource languages such as Luganda and Swahili may provide insights into processing and analysing those languages' textual data.

Data

- **What data will be used?**
  I will use the Sentiment Tagged Parallel Corpus for Luganda and Swahili [2], which contains parallel sentences in both languages tagged with sentiment labels.

- **How big is the data?**
  The corpus contains several thousand sentences in both languages. This provides sufficient data for clustering and analysis.

- **What are the attributes?**
  - Text in English: Sentences in English.
  - Text in Luganda: Translated sentences in Luganda.
  - Text in Swahili: Translated sentences in Swahili.
  - Sentiment Tags: Labels indicating sentiment (positive, negative).

Approach

- **What methods, algorithms, techniques will be used?**
  - **Text preprocessing:** This will include preprocessing such as tokenisation, lowercase, stop-word removal, stemming/lemmatization, and non-ASCII character handling.
  - **Feature extraction** using Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).
  - **Clustering techniques:** This will include using K-Means Clustering and Hierarchical Clustering to identify underlying patterns in data.
  - **Visualisation:** Word clouds will be used to highlight the most frequent words in each cluster to offer insights of each clusters. Additionally plots will be used to visualise the clusters.
  - **Topic Modeling:** Topic modelling methods, such as LDA, will be utilised to identify topics in the text. This will help in identifying hidden themes within the clusters and determining how these themes align with sentiment labels.

- **What are the expectations from these methods?**
  I expect to find unique clusters in the data that give insights into the textual data's structure, allowing us a better understanding of how clustering might be applied to low-resource languages. Additionally, I expect that topic modelling will reveal relationships between sentiment labels and theme topics. This could tell if specific themes in the corpus are more likely to express positive or negative emotion, providing an improved understanding of the data structure.

Evaluation

- **How will success be measured?**
  Success will be measured by evaluating cluster quality using metrics such as Silhouette Score. Word clouds will be used to visualise the most frequent topics within clusters. In addition, visual evaluations using t-SNE or PCA will be performed to determine clustering performance.

- **What are the baselines?**
  I will explore two clustering algorithms, K-Means and Hierarchical clustering. This comparison will help identify which method best represents the underlying data structure.

Expected Outcomes

- An evaluation of clustering techniques applied to sentiment data in Luganda and Swahili.
- Visualisations, such as cluster plots and word clouds, to highlight the most frequent words in each cluster.
- Insights into the effectiveness of feature extraction methods (BoW or TF-IDF) for clustering in low-resource languages.
- Possible recommendations will be given for future work on applying other clustering techniques to analyse sentiment data in low-resource languages.

References

[1] S. H. Muhammad, I. Abdulmumin, A. A. Ayele, *et al.*, *Afrisenti: A twitter sentiment analysis benchmark for african languages*, 2023. arXiv: 2302.08956 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2302.08956.

[2] C. Babirye, J. Tusubira, J. Mukiibi, J. Nakatumba-Nabende, and A. Katumba, *Sentiment Tagged Parallel Corpus for Luganda and Swahili*, version V1, 2023. DOI: 10.7910/DVN/XSGIKR. [Online]. Available: https://doi.org/10.7910/DVN/XSGIKR.