World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:17, No:12, 2023

The Advancements of Transformer Models in Part-of-Speech Tagging System for Low-Resource Tigrinya Language

Authors : Shamm Kidane, Ibrahim Abdella, Fitsum Gaim, Simon Mulugeta, Sirak Asmerom, Natnael Ambasager, Yoel Ghebrihiwot

Abstract : The call for natural language processing (NLP) systems for low-resource languages has become more apparent than ever in the past few years, with the arduous challenges still present in preparing such systems. This paper presents an improved dataset version of the Nagaoka Tigrinya Corpus for Parts-of-Speech (POS) classification system in the Tigrinya language. The size of the initial Nagaoka dataset was incremented, totaling the new tagged corpus to 118K tokens, which comprised the 12 basic POS annotations used previously. The additional content was also annotated manually in a stringent manner, followed similar rules to the former dataset and was formatted in CONLL format. The system made use of the novel approach in NLP tasks and use of the monolingually pre-trained TiELECTRA, TiBERT and TiRoBERTa transformer models. The highest achieved score is an impressive weighted F1-score of 94.2%, which surpassed the previous systems by a significant measure. The system will prove useful in the progress of NLP-related tasks for Tigrinya and similarly related low-resource languages with room for cross-referencing higher-resource languages.

Keywords : Tigrinya POS corpus, TiBERT, TiRoBERTa, conditional random fields
Conference Title : ICNLP 2023 : International Conference on Natural Language Processing
Conference Location : Istanbul, Türkiye
Conference Dates : December 18-19, 2023