PuoBERTa: Training and evaluation of a curated language model for Setswana

Vukosi Marivate[1,2][0000−0002−6731−6267], Moseli Mots'Oehli[3], Valencia Wagnerinst4[0000−0003−2671−7512], Richard Lastrucci[1], and Isheanesu Dzingirai[1]

[1] Department of Computer Science, University of Pretoria
vukosi.marivate@cs.up.ac.za
[2] Lelapa AI
[3] University of Hawaii at Manoa
moselim@hawaii.edu
[4] Sol Plaatje University
valencia.wagner@spu.ac.za

Abstract. Natural language processing (NLP) has made significant progress for well-resourced languages such as English but lagged behind for low-resource languages like Setswana. This paper addresses this gap by presenting PuoBERTa, a customised masked language model trained specifically for Setswana. We cover how we collected, curated, and prepared diverse monolingual texts to generate a high-quality corpus for PuoBERTa's training. Building upon previous efforts in creating monolingual resources for Setswana, we evaluated PuoBERTa across several NLP tasks, including part-of-speech (POS) tagging, named entity recognition (NER), and news categorisation. Additionally, we introduced a new Setswana news categorisation dataset and provided the initial benchmarks using PuoBERTa. Our work demonstrates the efficacy of PuoBERTa in fostering NLP capabilities for understudied languages like Setswana and paves the way for future research directions.

Keywords: Setswana, Natural Language Processing, Language Models