Abstract

Sentiment Analysis (SA) is a subset of Natural Language Processing (NLP) which has become a promising research area enabling the provision of language specific services. Although research in high resource languages such as English and Chinese has achieved promising results, research in low resource African languages such as Sesotho is still in its infancy due to limited text and speech datasets. This study contributes in this regard by availing the Sesotho News (SN) dataset, as an annotated dataset for the SA and Aspect Based Sentiment Analysis (ABSA) tasks. This dataset may be used for NLP research to benefit 1.85 million Sesotho speakers in Lesotho and 11.5 million speakers in South Africa. The dataset includes 4651 headlines for the ABSA task and 2401 headlines for the SA task using Lesotho's orthography of Sesotho. The news headlines were collected from Sesotho online newspapers and then annotated for the ABSA and SA tasks. The Spearman's correlation and Cohen's Kappa Index metrics show that there is good correlation between the annotators, implying that the SN dataset is of gold standard.