NATIVE TAMIL NEWS ARTICLE SUMMARIZATION

Satheeskumar.Y[1] and Chandana.G[2]

[1]*Department of Multimedia and Web Technology, University of Vocational Technology, Sri Lanka*
[2]*Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka*
mmw2014b106@uovt.ac.lk

Abstract: This research addresses the increasing trend of online news consumption within the Tamil-speaking community, driven by the growth of Internet technology. With the rise of online Tamil news platforms, readers can access a wide range of topics, but the absence of summarization in news articles poses challenges for efficient consumption. The research aims to employ machine learning techniques, specifically leveraging the XL-Sum model, for text summarization of Tamil news articles. The XL-Sum model comprises one million pairs of summaries and the news, annotated by professionals, covering 44 languages. Most research has focused on abstractive summarization in high-resource languages such as English, with limited work in low-resource languages like Tamil. The chosen XL-Sum model stands out as a large-scale multilingual abstractive summarization model and has been used with the custom dataset collected by the authors from selected Sri Lankan websites, consisting of 49 news and summary pairs annotated by professionals. The method involves gathering strategic data, selecting the XL-Sum pre-trained model, modifying and adapting it for Tamil, building a training dataset, re-training the model, and employing human evaluators to assess the quality of generated summaries, assigning ratings from high to medium. Key results include ROUGE scores comparing the new dataset with the existing model dataset, showing that the higher scores depend on the dataset's weight and fine-tuning the model. As an outcome, the authors can be able to release the new dataset for wider research community, providing news articles and their summarizations in Tamil from the Sri Lankan Community.

Keywords: Custom Dataset, Low-Resource Language Summarization, Machine Learning, Sri Lankan Tamil Dataset, XL-Sum Model.