ASPECTS IDENTIFICATION AND SENTIMENT ANALYSIS FOR SINHALA & ROMANIZED SINHALA SOCIAL MEDIA COMMENTS IN THE TELECOMMUNICATION DOMAIN

Kasun Lakshan Mahaliyanaarachchi

A dissertation submitted in partial fulfillment for the requirement for

Master of Science degree

in Big Data Analytics

Department of Computing

Informatics Institute of Technology, Sri Lanka

in Collaboration with

Robert Gordon University, Aberdeen, Scotland

2022

ABSTRACT

In the modern context of the business world, the customer experience department is vital in any kind of business. The profit of the company highly depends on the customer experience optimization strategies followed by the company. Therefore, implementing the best customer experience optimization strategies for the company is vital. Identifying the customer problems in real-time will help to improve the customer experience towards the brand. Social media is the best way to identify customer issues since people tend to express their feelings towards the company in social media as comments. Sentiment analysis and aspect predictions are done in this research to classify customer comments into different areas and to identify the sentiment of the comment. Research is done on the telecommunication domain since there is no such study done to the telecommunication domain previously and there is a high volume of data available in the social media compared to other domains. In the Sri Lankan context, most of the social media comments are based on the Singlish (Romanize Sinhala) and Sinhalese languages. Singlish (Romanize Sinhala) is the most commonly used method when writing comments on social media. Lack of Romanize Sinhala and Sinhala language resources has brought challenges from gathering and generating data sets to stemming, lemmatizing, and stop word removal. This research overcomes the above challenges by developing a Romanize Sinhala (Singlish) and Sinhala datasets for training the Aspect and sentiment prediction models and developing word embeddings for the both languages. Word2vec and FastText word embeddings are trained using Romanize Sinhala (Singlish) and Sinhala comments for the baseline model and identified the best word embedding model with the embedding size. Sentiment and aspect prediction models have trained afterward with the best word embedding model. The deep learning-based models such as GRU, LSTM, and CNN-based models were trained. All state-of-the-art models are outperformed by the proposed approach, which is based on capsule networks and the BI Directional GRU model. The accuracy, as well as weighted precision and recall, and weighted F1 scores, are used to determine which model is the most effective.

Key words: Sentiment Analysis, Capsule Network, BI Directional GRU