ASPECTS IDENTIFICATION AND SENTIMENT ANALYSIS FOR CODE-MIXED SINHALA-ENGLISH SOCIAL MEDIA COMMENTS IN THE TELECOMMUNICATION DOMAIN

N.A.H.W Shanaka Chathuranga

199309K

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

ASPECTS IDENTIFICATION AND SENTIMENT ANALYSIS FOR CODE-MIXED SINHALA-ENGLISH SOCIAL MEDIA COMMENTS IN THE TELECOMMUNICATION DOMAIN

N.A.H.W Shanaka Chathuranga

199309K

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2021

DECLARATION

I declare that this is my own work, and this thesis does not incorporate without acknowledgement any material previously submitted for a Postgraduate Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

UOM Verified Signature

The supervisor/s should certify the thesis/dissertation with the following Declaration.

The above candidate has carried out research for the Dissertation under my supervision.

Name of the supervisor: Dr. Surangika Ranathunga

Signature of the supervisor:

UOM Verified Signature

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 language. Singlish is the most commonly used method when writing comments on social media. Lack of Singlish 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 Singlish dataset for training the two models and developing word embeddings for the Singlish language. Word2vec and FastText word embeddings are trained using Singlish 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. Logistic regression, random forest, Naive Bayes, and SVM models are trained under the basic models.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

ACKNOWLEDGEMENT

I'd like to express my gratitude and appreciation to Dr. Surangika Ranathunga, Senior Lecturer, Department of Computer Science and Engineering, University of Moratuwa, for her guidance, assistance, and encouragement during this project. Without that direction, motivation, and help, this research would not have been possible. Also, I would like to thank Mr. Theekshana Madumal and Ms. Imali Limesha who supported me for the data annotation task.

I express my sincere gratitude to my parents, brother, sister, wife, and friends for their unwavering support throughout my life and for their contributions to my success and the success of this endeavor.

Finally, I want to express my gratitude to the person who helped me bring this study to fruition even if it was only with a word.

TABLE OF CONTENTS

LIST OF FIGURES