

**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

CHAPTER 1-INTRODUCTION	1
1.1 Introduction	1
1.2 Problem and Motivation	3
1.3 Objectives of Research	3
1.4 Organization of the Dissertation	3
CHAPTER 2- LITERATURE REVIEW	4
2.1 Introduction	4
2.2 Literature Review	4
2.2.1 Growth of Sentiment Analysis	4
2.2.2 Sentiment Analysis Methods	5
2.2.2.1 Lexicon-Based Method to do Sentiment Analysis	5
2.2.2.2 Machine Learning Approach for Sentiment Analysis	5
2.2.3 Sentiment Analysis using Deep Learning methods	6
2.2.3.1 Word Embeddings	6
2.2.3.2 Deep Learning Models	7
2.2.4 Sentiment Analysis for Indic Languages	8
2.2.4.1 Sentiment Analysis for the Sinhala Language	9
2.2.5 NLP approaches for Code-Mixed Data	9
2.2.6 Aspect Based Sentiment Analysis (ABSA)	11
2.2.6.1 Traditional techniques of ABSA	11
2.2.6.2 ABSA with Deep Learning	12
2.3 Discussion	13
CHAPTER 3 - METHODOLOGY	14
3.1 Introduction	14
3.2 High-Level methodology	14
3.3 Dataset creation	15
3.4 Aspect Definition	16
3.5 Data annotation	16
3.6 Data Cleaning and Pre-processing	18
3.7 Exploratory Data Analysis (EDA)	20
3.8 System Implementation	21

3.8.1 Word Embeddings	21
3.8.2 Classification model Training	21
3.9 Output	26
3.10 Summary	26
CHAPTER 4-EVALUATION	27
4.1 Introduction	27
4.2 Exploratory Data Analysis (EDA)	27
4.2.1 Number of comments	27
4.2.2 Class distribution of sentiment	28
4.2.3 Class imbalance in the Sentiment analysis model	29
4.2.4 Distribution of comment length	29
4.2.5 Number of words Distribution	31
4.2.6 Most common words according to the sentiment	33
4.2.7 Distribution of Singlish words percentage	35
4.2.8 Distribution of aspects	36
4.2.10 Distribution of multiple tags per comment	38
4.2.11 Words distribution according to the aspect	38
4.2.12 Word2vec word vectors	41
4.2.13 Inter-Annotator Agreement score	42
4.3 Classification results	42
4.3.1 Sentiment Analysis model results -Basic Machine Learning models	43
4.3.1.1 Effectiveness of the text cleaning	46
4.3.2 Sentiment Analysis model results -Deep Learning models	47
4.3.2.1 Deep Learning models	47
4.3.3 Aspect prediction model results -Basic Machine Learning models	48
4.3.4 Aspect prediction model results -Deep Learning models	51
4.3.5 Hyper-Parameter tuning results	51
4.3.5.1 Hyper-Parameter tuning results - Sentiment analysis	52
4.3.5.1.1 Learning curves	53
4.3.5.2 Hyper-Parameter tuning results -Aspect prediction model	55
4.3.5.2.1 Aspect prediction model results -Learning curves	56
CHAPTER 5 -DISCUSSION	59
CHAPTER 6 – CONCLUSION AND FUTURE WORKS	61

References

63

Appendix

69

LIST OF FIGURES

Figure 3. 1 High Level Architecture	15
Figure 3. 2 Proposed Model.....	23
Figure 3. 3 Capsule Network [46].....	25
Figure 4. 1 Results of the study	27
Figure 4. 2 Class Distribution of Sentiment.....	28
Figure 4. 3 Distribution of Comment Length	29
Figure 4. 4 Distribution of comment length with Sentiment	30
Figure 4. 5 Distribution of Number of Words	31
Figure 4. 6 Distribution of Number of Words with Sentiment	32
Figure 4. 7 Positive Common Words.....	33
Figure 4. 8 Negative Common Words	34
Figure 4. 9 Neutral Common Words.....	34
Figure 4. 10 Distribution of Single Words Percentage	35
Figure 4. 11 Aspect Distribution.....	36
Figure 4. 12 Aspect Distribution According to Sentiment.....	37
Figure 4. 13 Distribution of multiple tags per comment.....	38
Figure 4. 14 Word distribution of Network Aspect	38
Figure 4. 15 Word distribution of Billing/Price Aspect.....	39
Figure 4. 16 Word distribution of Package Aspect.....	39
Figure 4. 17 Word distribution of Customer Service Aspect.....	40
Figure 4. 18 Word distribution of Data Aspect.....	40
Figure 4. 19 Word distribution of Service/Product Aspect.....	41
Figure 4. 20 Multiclass ROC Curve.....	46
Figure 4. 21 Sentiment Model: Training and Validation Loss.....	53
Figure 4. 22 Sentiment Model: Training and Validation Accuracy.....	54
Figure 4. 23 Sentiment Model: Training and Validation F1- Score	54
Figure 4. 24 Aspect Model: Training and Validation Loss.....	57
Figure 4. 25 Aspect Model: Training and Validation Accuracy	57
Figure 4. 26 Aspect Model: Training and Validation F1 Score	58