

MSc In Information Technology

**Customer Satisfaction Monitoring with Sentiment Analysis Based on
Twitter Feeds in Telecom Domain**

Prepared by

A.P.L.D. Sachith Chamara

(158752P)

Supervised by

Mr. S.C. Premaratne

Faculty of Information Technology

University of Moratuwa

2018

MSc In Information Technology

**Customer Satisfaction Monitoring with Sentiment Analysis Based on
Twitter Feeds in Telecom Domain**

Prepared by
A.P.L.D. Sachith Chamara
(158752P)

Supervised by
Mr. S.C. Premaratne

Faculty of Information Technology
University of Moratuwa

2018

Declaration

I declare that this thesis/dissertation does not incorporate without acknowledgement any material, previously submitted for a Degree or Diploma in any University or other 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.

A.P.L.D. Sachith Chamara
(158752P)
2018/05

I have supervised and accepted this thesis for the submission of the degree.

Mr. S.C. Premaratne
(Main Supervisor)
2018/05

Acknowledgements

I would like to take this opportunity to thank my supervisor, Mr. S.C. Premaratne for all his dedicated support and guidance throughout this research. His guidance and inspiration are the key factors that enabled the successful completion of this research.

I would like to express my sincere thanks to Mr. Ian Ramsden (F IDM) for providing an excellent introduction to this research area. His valuable discussions and comments were also effective stimuli to this research work.

My deepest thanks go to my family for their assistance, encouragement and unwavering belief in me. Finally, I would like to thank all the staff of University of Moratuwa who helped me directly or indirectly to carry out this research work.

Abstract

With this increased competition among telecom service providers, it has become more difficult to retain the existing customers, but when the number of customers reaches its peak, finding and securing new customers become increasingly difficult and costly. Therefore, it would be better to prioritize the retention of the existing customers, than trying to win new ones.

Customer reviews can be recognized as fruitful information sources for monitoring and enhancing customer satisfaction levels as they convey the real voices of actual customers expressing relatively unambiguous opinions.

This research is aimed at mining and measure customer satisfaction toward Telecom Service based on reviews and feedbacks from Twitter. This research is mainly focus on one of the largest mobile operator in Sri Lanka and the analysis has been done only for English language.

Tweets were classified into three classes as Positive, Negative and Neutral with the use of four dictionaries (Lexicon, SentiWordNet, Slangs& Emoticons). The framework was built based on six steps and it shows that Lexicon performs well on the dataset better than SentiWordNet. After fine-tuning lexicon and stop words dictionary and integrating with Slangs dictionary, positive classification shows 91.98% accuracy without Emoticon dictionary while for negative classification, the accuracy is 82.27% with Emoticons dictionary.

Keywords: Twitter Feeds, Telecom Industry, Sentiment Analysis, Lexicon

Table of Contents

Declaration	i
Acknowledgements	ii
Abstract	iii
List of Figures	vi
List of Tables	vi
1 Introduction	1
1.1 Prolegomena	1
1.2 Problem Statement	1
1.3 Aims & Objectives	1
1.4 Background and Motivation	2
1.5 Problem in brief	3
1.6 Proposed Solution	4
1.7 Structure of the thesis	4
2 Literature Review	5
2.1 Introduction	5
2.2 Customer Satisfaction Monitoring Framework	5
2.3 Goal Setting	5
2.4 Text Preprocessing	5
2.5 Parsing the content	6
2.6 Text Refinement	7
2.7 Analyzing and Scoring	7
2.7.1 Analyzing	7
2.7.2 Scoring	10
2.8 Finalize and validate the model	10
2.8.1 The true positive rate (Hit rate or Recall) of a classifier	11
2.8.2 The false positive rate of a classifier (FPR)	11
2.8.3 Precision	12
2.8.4 F-Score	12
2.9 Summary	12
3 Technology	13
3.1 Introduction	13
3.2 Machine Learning Vs Lexicon Based Method	13
3.3 Limitations of Lexicon Based Method	13
3.4 Summary	14
4 Approach	15
4.1 Introduction	15
4.2 Hypothesis	15
4.3 Input	15
4.4 Output	16
4.5 Process	16
4.6 Summary	16
5 Analysis & Design	17
5.1 Introduction	17
5.2 High-Level Solution Diagram	17
5.2.1 Text Preprocessing Module	18
5.2.2 Content Parsing Module	19
5.2.3 Lexicon Module	19

5.2.4 Scoring Module.....	20
5.2.6 Finalize and validate the model	21
5.3 Summary	22
6 Implementation	23
6.1 Introduction	23
6.2 Approach	23
6.3 Dataset.....	23
6.4 Text Preprocessing Module	23
6.5 Content Parsing Module.....	24
6.6 Lexicon Module	24
6.7 Scoring Module.....	24
6.7.1 Algorithm for using Lexicon Dictionary only.....	24
6.7.2 Algorithm for using SentiWordNet only	25
6.7.3 Algorithm for using both Lexicon and SentiWordNet.....	26
6.7.4 Algorithm for using Lexicon with Slang Replacements.....	26
6.7.5 Algorithm for using Lexicon with Slang Replacements and Emoticons Dictionary.....	27
6.8 Summary	28
7 Discussion	29
7.1 Introduction	29
7.2 Results & Analysis	29
7.3 Model Validation with existing tools/APIs.....	34
7.3.1 Google Cloud Natural Language API.....	34
7.3.2 tap.aylien.com API	35
7.3 Research Limitations	36
7.4 Suggestions for Further Research	37
7.5 Summary	37
References	38
Appendixes	40
<i>Appendix A – R code for using only Lexicon Dictionary</i>	<i>40</i>
<i>Appendix B – R code for using SentiWordNet</i>	<i>44</i>
<i>Appendix C – R code for using Lexicon Dictionary with Stop Words Amendments</i>	<i>48</i>
<i>Appendix D – R code for comparing Lexicon Dictionary with SentiWordNet</i>	<i>51</i>
<i>Appendix E – R code for using Lexicon Dictionary with Slang Replacements</i>	<i>55</i>
<i>Appendix F – R code for using Lexicon Dictionary with Slang Replacements and Emoticons.....</i>	<i>59</i>

List of Figures

Figure 1.1: Cellular Mobile Telephone Subscriptions [1]	2
Figure 2.1: Stemming Process [9].....	6
Figure 2.2: Term-By-Document Matrix [11]	7
Figure 2.3: Sentiment Classification Techniques [15].....	9
Figure 2.4: Sample score for sentiment [6].....	10
Figure 2.5: Confusion matrix for binary classification [18].....	11
Figure 5.1: High Level Solution Diagram.....	17
Figure 7.1: Comparison of Positive Classification Accuracy for each step	33
Figure 7.2: Comparison of Negative Classification Accuracy for each step	34
Figure 7.3: Snapshot of Sentiment Score on Google Cloud API.....	35
Figure 7.4: Snapshot of tap.aylien.com API.....	36

List of Tables

Table 5.1: Confusion Matrix for Positive Reviews	21
Table 5.2: Confusion Matrix for Negative Reviews	21
Table 5.3: Confusion Matrix for Neutral Reviews	22
Table 6.1: Main Approaches	23
Table 7.1: Lexicon vs SentiWordNet	29
Table 7.2: Samples of Fine-tuned Words	30
Table 7.3: Lexicon vs Fine-tuned Lexicon.....	30
Table 7.4: Fine-tuned Lexicon vs Fine-tuned Lexicon with Stop Words Changes.....	31
Table 7.5: Hybrid Approach (Both Lexicon and SentiWordNet).....	31
Table 7.6: Lexicon vs Lexicon with Slang Replacements.....	32
Table 7.7: Lexicon with Slang Replacements and Emoticons	33
Table 7.8: Accuracy Comparison with Google Cloud API.....	35
Table 7.9: Accuracy comparison with tap.aylien.com API.....	36