MSc in Information Technology

Analyze Quality of Products in E-commerce Systems with Sentimental Analysis

Prepared by

P. M. A. U. Bandara

Index No : 158751L

Supervised by

Mr. SamindaPremaratne

Faculty of Information Technology

University of Moratuwa

March 2019

MSc in Information Technology

Analyze Quality of Products in E-commerce Systems with Sentimental Analysis

Prepared by

P. M. A. U. Bandara

Index No: 158751L

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology.

March 2019

Declaration

I declare that this research is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and the list of references is given.

P. M. A. U. Bandara

(158751L)

2019/04/

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

Mr. Saminda Premaratne

(Main Supervisor)

2019/04/

Acknowledgements

I would like to express my genuine gratitude to my advisor Mr. S.C Premaratne for the continuous support of my research, for his patience, motivation, and immense knowledge. His direction helped me in all the period of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my research. Also I thank my associates and staff in the University of Moratuwa who helped me carry out this research. Last but not the least, I would like to thank my family, my parents for supporting me spiritually throughout writing this thesis.

Abstract

E-commerce websites getting extra significant and popular today since the vast differentiated and diversified information that is presented. Studies says that more than 80% of the world population is using these websites to purchase goods and services online. For these online customers, comments / feedbacks play a major role in decision making when buying the products from the market space. Hence the diversity and the popularity of the Online space, sales of these online products get increased with time. Therefore, it is not practical to review all the given product feedback and come to conclusion on purchasing the product for a consumer. Focusing on this, this study is urges to observe the success factors of online websites and how those aspects influence on online marketing to sales growth in any organization. Therefore, in this research a Analyze Quality of Products in E-commerce System is focused on analyzing the online consumers feedbacks or comments on various products using data mining techniques such as Sentimental and filtering analysis. The outcome from the study will show feature wise relativeness in the mobile phone domain. All procedures were based on the features extracted through a thorough literature review and existing apparatuses. This will aid to calculate a “Trust Score” for the online products and a general overview to achieve a higher trust score for e-commerce organization.

Table of Contents

Declaration I

Acknowledgements II

Abstract III

Table of Contents IV

List of Figures VII

List of Tables VII

Chapter 1 – Introduction 1

1.1 Prolegomena 1

1.2 Identifying the Research Problem 2

1.2.1 Symptoms of the Problem 2

1.2.2 Justification of the problem 3

1.2.3 Defining Research Problem 3

1.3 Research Question 5

1.4 Aim 5

1.5 Objectives of the Study 5

1.6 Study Limitations 5

1.7 Chapter Outline 5

1.8 Chapter Summary 6

Chapter 2 – Literature Review on Developments and Challenges in Product Quality Analysis 7

2.1 Introduction 7

2.2 Early development and Modern Trends 7

2.3 Existing Models 8

2.4 Lexicon-Based Approach 8

2.5 Confusion Matrix 9

2.6 Summary 12

Chapter 3 – Adopted Technology in Data Mining Tools and Techniques 13

3.1 Introduction 13

3.2 Comparison of Machine Learning and Lexicon Based Method 13

3.3 What is Data Mining? 14
3.4 R Data Mining Tool 15
3.5 Summary 15

Chapter 4 - Approach to Analyze Quality of Products in a System 16
4.1 Introduction 16
4.2 Users 16
4.3 Input 16
4.4 Output 16
4.5 Process 16
4.6 Technology 17
4.7 Summary 17

Chapter 5 – Design to Analyze Quality of Products in System 18
5.1 Introduction 18
5.2 Top level Architecture of the System 18
5.3 Data Collection 19
5.4 Pre-Processing 19
5.4.1 Removal of duplicate 20
5.4.2 Removal of HTML tags , hyperlinks 20
5.4.3 Stop words removal 20
5.4.4 Tagging Speech 20
5.5 Tokenization 21
5.6 Categorization 21
5.7 Lexicon Module 21
5.7.1 Lexicon Dictionary 21
5.8 Generating Product Rating 22
5.9 Rating Module 22
5.10 Validate Model 22
5.10.1 Confusion matrix for positive reviews 23
5.10.2 Confusion matrix for negative reviews 23
5.10.3 Confusion matrix for neutral reviews 23
5.11 Summary 24

Chapter 6 – Implementation of Analyzing the Quality of the Product in a System 25
6.1 Introduction 25
6.2 R Language 25
6.3 Data Collection 26
6.4 Data set Preprocessing 26
6.5 Categorization of the Dataset 26
6.6 Implementation of the Application 26
6.6.1 Lexicon Module 26
6.6.2 Lexicon Dictionary 27
6.6.3 Scoring Algorithm 27
6.7 Summary 27
Chapter 7 - Evaluation 28
7.1 Introduction 28
7.2 Analysis 28
7.3 Model Validation with Existing Products 31
7.3.1 TheySay API 31
7.3.2 Google Cloud Natural Language 32
7.3.3 Paralleldots API 33
7.4 Overview of the Research 34
Chapter 8 – Conclusion and Recommendations 35
8.1 Introduction 35
8.2 Summary of the research 35
8.3 Conclusions 35
8.4 Limitations of the Study 35
8.5 Future Work 36
Appendixes 41
Appendix A – R code for Web Scrape from Amazon 41
Appendix B – R code for Web Scrape Parse Amazon html pages for data 42
Appendix C – R code for Review categorized by features 46
Appendix D – R code for Sentimental Analysis using Lexicon Dictionary 48

List of Figures

Figure 1.1.1: Retail e-commerce sales worldwide from 2014 to 2021 1

Figure 2.4.1: Example of Native words 9

Figure 3.3.1: Steps of data mining Process 14

Figure 5.2.1: Top level architecture of the system 19

Figure 7.2.1: Positive Accuracy Comparison for Lexicon 30

Figure 7.2.2: Negative Accuracy Comparison for Lexicon 30

Figure 7.3.1: Result of theysay API 31

Figure 7.3.2: Google Cloud Natural Language Results 32

Figure 7.3.3: Google Cloud Natural Language Results 33

Figure 7.3.4: Paralleldots Results 33

List of Tables

Table 5.11.1: Confusion Matrix for Positive Reviews 23

Table 5.11.2: Confusion Matrix for Negative Reviews 23

Table 5.11.3: Confusion Matrix for Neutral Reviews 23

Table 7.2.1: Lexicon Performance 28

Table 7.2.2: Tweaked for Mobile Phone Domain 29

Table 7.2.3: Lexicon VS Tweaked Lexicon 29

Table 7.3.1: Comparison with TheySay API 31

Table 7.3.2: Comparison with Google Cloud Natural Language 33

Table 7.3.3: Comparison with Paralleldots API 34