Data Extraction from Social Media for Sentiment Analysis in order to Predict Sales and Correlated Items for Fashion Industry

C.S Kumanayaka

169315X

Faculty of Information Technology

University of Moratuwa

December 2018

Data Extraction from Social Media for Sentiment Analysis in order to Predict Sales and Correlated Items for Fashion Industry

C.S Kumanayaka

169315X

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

December 2018

Declaration

I declared that this thesis is written by myself and it has not submitted by other institution of education, degree or diploma of any other university. Information retrieved from unpublished and published work listed in the reference area.

Name of Student

Signature of Student

Date:

Supervised by

Name of Supervisor(s)

Signature of Supervisor(s)

Date:

Acknowledgment

Initially I am thankful to my supervisor Mr. Saminda Premaratne the valuable advice and supervision. I like to give my gratitude to him for his important time for long discussions throughout my research. His rightful guidance paved the way to achieve priceless achievements

Special thanks should also go to Mr Sankha Perera for the advice and guidance provided from a Statistical point of view. I am also thankful for Mr Prabhath Gunawardane for his valuable guidance and advice in Data Science. Gratitude should also go to Ms.Ayesha Kasthuriarachchi for the statistical advice and guidance.

Also, I should be thankful to all the lecturers about the assistance and guidance provided in carrying out the project. Finally, I should also be grateful to all my batch mates and all the other parties who helped us with their valuable ideas to proceed with my research successfully.

Abstract

These days' Social media is being very popular for marketing. One of the most popular social media platforms is Facebook. Most of the Fashion brand have Facebook pages. Consumer expresses their feeling using comments and emotional buttons. The user interacts with the brand page using post, like, share comments. The analyzed data can give support to decision makers to the evaluation of the customer's feedback, identify potential customers and predict sales item for the upcoming month and predict correlated items. The purpose of this paper is to explain how to extract and prepare data collected on Facebook to perform sentiment Analysis.

Table of Content

Declaration ii

Acknowledgement iii

Abstract iv

Table of Content v

List of Figures vi

Abbreviations viii

Chapter 1 1

Introduction 1

SQL Server Data Tools (SSDT) for Visual Studio 4

Chapter 2 5

Literature Review 5

Chapter 3 13

Technologies Used for the Tool 13

Chapter 4 16

Analysis and Design 16

4.1 Social media data extraction 17

4.2 Preprocessing of Text data 17

4.3 Feature Extraction using Text Data 17

4.5 Sentiment Analysis (Sentiment model building and training) 18

4.6 Predict Sales using Sentiment and Semantic (Sales model building and training) 19

4.7 Test and evaluation 19

4.8 Identify correlated item 20

4.9 Live Dashboard (Correlated item, sales prediction and insight about items) and mobile application 20

Chapter 5 21

Implementation 21

5.1 Data Collection 21

5.2 Data Preprocessing 22

5.3 Feature Extraction using Text Data 28

5.4 Split training and testing data 32

5.5 Sentiment Analysis (Sentiment model building and training)............................33
5.6 Predict Sales using Sentiment and Semantic (Sales model building and training) 36
5.7 Predicting Sales item and correlated item 38
5.8 Live Dashboard (Correlated item, sales prediction and insight about items) and mobile application 39
Chapter 6...41
Evaluation 41
6.1 Accuracy Check the textblob library...41
6.2 Accuracy of Sentiment model 41
6.3 Accuracy of Sales model..42
Chapter 7...43
Conclusion and further work 43
7. 1 Overall Achievement..43
7. 2 Achievement of each objective...43
7. 3 Problem encountered..44
7.4 Limitation 44
7.5 Further work 44
References..45

List of Figures

Figure1.1 Research overview diagram Terminology 04

Figure 2.1 Facebook Terminology [2] 08

Figure 2.2 Part of the dictionary of emoticons and emoji’s [2] 08

Figure 2.3 System Design 09

Figure 2.4 Performance Evaluation 11

Figure 2.5 Social Media is important to my business 12

Figure 2.6 Weekly time commitment for social media marketing 12

Figure 4.1 System design 17

Figure 4.2 Train/Text Split 19

Figure 5.1 Raw data from ‘Thefashionworks’ facebook page 22

Figure 5.2. Data cleaning 23

Figure 5.3 Lower casing 24

Figure 5.3 Remove Punctuation 24

Figure 5.4 Remove Stopwords 25

Figure 5.5 Remove Common Word 26

Figure 5.6 Remove Rare Words 27

Figure 5.7 Spelling Correction 27

Figure 5.8 Tokenization 28

Figure 5.9 Stemming 28

Figure 5.10 Lemmatization 29

Figure 5.11 Number of words in the comments 30

Figure 5.12 Number of characters in the comments 30

Figure 5.13 Average word length in the comments 31

Figure 5.14 count number of stop word from comments 31

Figure 5.15 Number of special character word from comments 32

Figure 5.16 generate number of numeric from comments 32

Figure 5.17 Generate number of upper case word in the comments 33

Figure 5.18 splitting the training and testing data set. 34

Figure 5.19 detecting the sentiment using the textblob library 34

Figure 5.20 python code for splitting and setting dependent variable 35

Figure 5.21 the python code for feature matric creation. 35

Figure 5.22 python code fitting and training model 36

Figure 5.23 Predict the test data using model 36

Figure 5.24 plotting the model 36

Figure 5.25 plot sentiment model 37

Figure 5.26 Python scripts for dependent/independent variable setting and fitting the model 38

Figure 5.27 some of sales prediction for testing data set 38

Figure 5.28 plotting the sales model 39

Figure 5.29 exiting item category 40

Figure 5.30 check correlation - python script and results. 41

Figure 5.31 Some charts of dashboard 42

Figure 6.1 Checking Accuracy 43

Figure 6.2 Shows the python scrip for generate accuracy of sales model 44

Abbreviations

FB	Facebook
ETL	Extraction transformation load
SSIS	SQL Server Integration service
SSRS	SQL Server Reporting service
BI	Business Intelligence