SOCIAL MEDIA SENTIMENT ANALYSIS BASED ON AFFECTIVE-BEHAVIOURAL-COGINITIVE MODEL OF ATTITUDES

Dewamuni Adikaramge Chamodi Madhushani

(179333J)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2020

SOCIAL MEDIA SENTIMENT ANALYSIS BASED ON AFFECTIVE-BEHAVIOURAL-COGINITIVE MODEL OF ATTITUDES

Dewamuni Adikaramge Chamodi Madhushani

(179333J)

Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science Specializing in Data Science, Analytics and Engineering

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

May 2020

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of Higher Learning and to the best of my knowledge and belief this 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 dissertation, 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).

Name: D. A. C. Madhushani

Signature: …….. Date:.....

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

Name of the Supervisor: Dr. Surangika Ranathunga

Signature: ……. Date:.....

ABSTRACT

Sentiment Analysis is the study of classifying a given text based on its sentiment (positive/ negative polarity) of the expression. Sentiment analysis is being widely used to analyse the public opinion towards a given entity. Today in Web 2.0, social media is a popular platform to express one's opinions and beliefs. Therefore, researchers are keen on investigating how social media sentiment analysis can be improved to benefit interested entities. Most of the sentiment analysis research has been conducted on identifying the polarity (i.e.: positive, negative or neutral) and emotions (i.e.: happiness, sadness, disgust, anger, fear and surprise).

Comparatively, less focus has been given to study how expressions can be classified based on psychological aspects of attitude. The objective of the proposed research is to move beyond the mere polarity, and to investigate whether we can get an in-depth understanding of the expressed attitude. For this, we have used the ABC (Affective, Behavioural and Cognitive) model of attitude introduced in consumer psychology.

In this research a new dataset was compiled by extracting Tweets on a specific topic and manually annotating them based on the attitude by domain experts. This research discusses how existing tools and technologies of Sentiment Analysis can be applied for this problem domain. Various preprocessing and feature extraction techniques were evaluated against a set of machine learning algorithms including Ensemble and Deep Learning models. Additionally, this research aims to contribute to reduce the gap between machine learning and consumer psychology and thereby proving the possibility of applying machine learning across different domains.

DEDICATION

I dedicate this Masters dissertation to my beloved parents, Mrs. Rupika Wijesekara and Mr. Shantha Adikaram. I hope that this achievement will be one step closer to completing the dream that you had for me all those many years ago when you chose to give me the best education you could.

ACKNOWLDGEMENT

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Surangika Ranathunga for providing abundant guidance, support and encouragement throughout this research. Special thanks to Ms. Irosha Amarabandu, Research Officer, Department of Psychiatry, Faculty of Medical Sciences, University of Sri Jayewardenepura for giving her invaluable support to conduct this research. Also, I would like to thank my colleagues for sharing knowledge, support and constant encouragement. I express my gratitude to University of Moratuwa for providing me with necessary resources and knowledge to make this project a success.

Last but not least, I express my love and gratitude to my beloved family for their continuous and unparalleled love, support and understanding.

TABLE OF CONTENTS

DECLARATION i
ABSTRACT ii
DEDICATION iii
ACKNOWLDGEMENT iv
TABLE OF CONTENTS v
LIST OF FIGURES viii
LIST OF TABLES ix
LIST OF EQUATIONS xi
LIST OF ABBREVIATIONS xii
1. INTRODUCTION 1
1.1 Sentiment Analysis 1
1.2 Types of attitudes 2
1.2.1 Tricomponent Attitude Model (ABC model of attitude) 3
1.3 Problem Statement and Motivation 4
1.4 Objectives 6
1.5 Research Scope 7
1.6 Thesis Organization 7
2. LITERATURE REVIEW 8
2.1 Pre-processing 9
2.1.1 Handling Numbers 9
2.1.2 Handling Punctuations 9
2.1.3 Lowercasing 10
2.1.4 Replace Slang and Abbreviations 10
2.1.5 Spelling Correction 10
2.1.6 Handling Negations 11
2.1.7 Replace URLs, Hashtags and User Mentions 11
2.1.8 Handling Emoticons 12
2.1.9 Tokenization 12
2.1.10 Part of Speech Tagging 12
2.1.11 Stop words removal 12
2.1.12 Stemming and Lemmatization 13
2.2 Feature Extraction 14
2.2.1 Bag of words 14

2.2.2 Word N-grams 14
2.2.3 Word Embedding 14
2.2.4 Word Clusters 15
2.2.5 Part-of-speech Features 15
2.2.6 Emoticons 15
2.2.7 TF-IDF 15
2.2.8 Statistical and Meta Information on a Message 15
2.3 Feature Selection 16
2.3.1 Frequency cut off 16
2.3.2 HMM LSI and LDA 16
2.3.3 Pointwise Mutual Information 17
2.3.4 Chi-square 17
2.4 Sentiment Classification 18
2.4.1 Lexicon Based Approach (LB) 18
2.4.1.1 Dictionary Based Approach 19
2.4.1.2 Corpus Based Approach 19
2.4.2 Machine Learning Approach (ML) 20
2.4.2.1 Supervised Learning 20
2.4.2.2 Unsupervised Learning 23
2.4.3 Ensemble Learning (EL) 23
2.5 Evaluation 25
2.6 Types of Sentiment Classification 26
2.7 Applications of ABC Model of Attitudes 29
3. RESEARCH METHODOLOGY 31
3.1 Solution Architecture 31
3.2 Data Collection 31
3.2.1 Data Extraction 33
3.2.2 Data Annotation 35
3.3 Exploratory Data Analysis (EDA) 37
3.4 Implementation of Preprocessing 42
3.5 Implementation of Feature Extraction 49
3.6 Implementation of Feature Selection 51
3.7 Implementation of Classification 51
3.7.1 Multinomial Naive Bayes (MNB) 52
3.7.2 Logistic Regression 52

3.7.3 Support Vector Machines (SVM) 53
3.7.4 Random Forest 53
3.7.5 Ensemble Classifiers 54
3.7.6 Simple Neural Network 55
3.7.7 Convolutional Neural Network (CNN) 55
3.7.8 Recurrent Neural Network (RNN) 57
3.8 Handling Imbalanced Class Problem 58
3.8.1 Re-sampling 58
3.8.2 Class Weights 59
3.8.3 Hyper Parameter Optimization 59
3.9 Evaluation Metrics 60
4. SYSTEM EVALUATION 61
4.1 Inter-Rater Agreement 61
4.2 Baseline Experiment 63
4.3 Effect of Preprocessing Techniques 64
4.4 Effect of Resampling 67
4.5 Effect of Feature Selection 68
4.6 Evaluation of Models 69
4.6.1 Word Embeddings 75
4.6.2. Deep Learning Models 77
4.7 Error Analysis 79
4.8 Summary 80
5. CONCLUSION 83
5.1 Future Improvements 84
REFERENCES 85

LIST OF FIGURES

Figure 1.1: What affects the customers buying behaviours and choices 5
Figure 1.2: Affective advertising 5
Figure 1.3: Behavioural advertising 6
Figure 1.4: Cognitive advertising 6
Figure 2.1: Steps of SA Model 8
Figure 2.2: Classification of SA techniques 18
Figure 2.3: Plutchik's Wheel of Emotions 27
Figure 3.1: Code Snippet to Include Emojis in Text 34
Figure 3.2: Sample of annotated dataset 36
Figure 3.3: Distribution of Classes 37
Figure 3.4: Length Distribution of Tweets 38
Figure 3.5: Top 50 Most frequent terms before preprocessing 39
Figure 3.6: Top 50 Most frequent terms after preprocessing 39
Figure 3.7: Top 50 most occurring words in Affective class 40
Figure 3.8: Top 50 most occurring words in Behavioural class 41
Figure 3.9: Top 50 most occurring words in Cognitive class 41
Figure 3.10: Process flow- Preprocessing 42
Figure 3.11: CNN architecture 56
Figure 3.12: RNN-LSTM architecture 57
Figure 4.1: Feature Importance Graph (Logistic Regression with TF-IDF features) 70

LIST OF TABLES

Table 1.1: Example of classification of attitudes based on ABC model 4
Table 2.1: Confusion Matrix..... 25
Table 3.1: Categorization and Distribution of Classes 37
Table 3.2: Example- Handling Twitter tags..... 43
Table 3.3: Example- Handling Emojis 44
Table 3.4: Example- Lowercasing..... 45
Table 3.5: Evaluation of Different Spell Correctors..... 46
Table 3.6: Example- Handling Numbers 47
Table 3.7: Example- Handling Punctuations 47
Table 3.8: Example- Handling Negation 48
Table 3.9: Selected Supervised Learning Classifiers..... 51
Table 4.1: Inter Rater Agreement 61
Table 4.2: Results of Baseline Experiment..... 64
Table 4.3: Effect of Preprocessing Techniques 65
Table 4.4: Results of Preprocessing Step Vs Baseline Model..... 67
Table 4.5: Effect of resampling techniques 68
Table 4.6: Effect of Feature Selection 68
Table 4.7: Results of SVM Model..... 69
Table 4.8: Results of Logistic Regression Model..... 70
Table 4.9: Baseline and Optimized Hyper Parameter values- Logistic Regression .. 71
Table 4.10: Effect of Parameter Tuning on Logistic Regression with TF-IDF Features..... 71
Table 4.11: Effect of Balanced Class weights on Logistic Regression with TF-IDF Features..... 72
Table 4.12: Confusion Matrix Logistic Regression with TF-IDF Features (class_weights= None)..... 72
Table 4.13: Confusion Matrix Logistic Regression with TF-IDF Features (class_weights= ‘balanced)..... 72
Table 4.14: Results of Multinomial Naive Bayes, Random Forest, Stacking, Random Subspace and Simple Neural Network 74

Table 4.15: Results of Using Word Embedding features with SVM and Logistic Regression Classifiers 75
Table 4.16: Results of FastText Text Classification (Multinomial Logistic Regression) 76
Table 4.17: Results of Deep Learning Model CNN and RNN-LSTM using Word Embedding features 77
Table 4.18: Classification report Logistic Regression Vs CNN 78
Table 4.19: Error Analysis 79
Table 4.20 Summary of Best Results 80
Table 4.21: Hyper-parameters of Best Performing Models 82

LIST OF EQUATIONS

Equation 2.1: Pointwise Mutual Information 17
Equation 2.2: Chi-Square..... 17
Equation 2.3: Pobability Calculation of Maximum Entropy 21
Equation 2.4: Accuracy..... 25
Equation 2.5: Precision..... 25
Equation 2.6: Recall..... 25
Equation 2.7: F1 Score 25
Equation 4.1: Cohen's Kappa Score 62

LIST OF ABBREVIATIONS

ABC	Affective Behavioural Cognitive
BOW	Bag of Words
CBOW	Continuous Bag of Words
CNN	Convolution Neural Networks
EDA	Exploratory Data Analysis
FS	Feature Selection
HMM	Hidden Markov Model
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
POS	Part of Speech
RNN	Recurrent Neural Networks
ROS	Random Over-sampling
SA	Sentiment Analysis
SC	Sentiment Classification
SMOTE	Synthetic Minority Oversampling
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
TPE	Tree of Parzen Estimators