



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم



شبكة المعلومات الجامعية التوثيق الإلكتروني والميكروفيلم



MONA MAGHRABY



شبكة المعلومات الجامعية
التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغييرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



MONA MAGHRABY



**FACULTY OF COMPUTER
& INFORMATION SCIENCES
AIN SHAMS UNIVERSITY
Abbassia, Cairo, Egypt**

Identifying User Attitude Using Sentiment Analysis in Social Media

A Thesis submitted to Computer Science Department, Faculty of
Computer and Information Sciences, Ain Shams University, in partial
fulfillment of the requirements for the degree of Doctor of
Philosophy in Computer Science

BY

Soha Saied Ibrahim ElShafie

Faculty of Computer and Information Sciences Ain Shams University

Under the Supervision of

Prof. Dr. Mostafa Mahmoud Aref

Computer Science Department,
Faculty of Computer and Information Sciences
Ain Shams University

Prof. Dr. Khaled Abdel Hamid Bahnasy

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Dr. Sally Ismail Saad

Computer Science Department,
Faculty of Computer and Information Sciences
Ain Shams University

2020

Acknowledgement

All praise and thanks to ALLAH, who provided me with the power to complete this work. I hope from ALLAH to accept this work from me.

I am glad to present my sincere appreciation to my parents who provided me with their guidance, help and support. I would like to give my husband a special “Thank You!” as he encouraged and supported me a lot.

I also present my sincerest gratitude to my supervisors, Prof. Dr. Mostafa Aref and Dr. Khaled Abdel Hamid Bahnasy and Dr. Sally Saad who have supported me throughout my thesis with their patience, knowledge and experience.

Last but not least, I would like to thank my sisters, my friends, and all my family who supported and encouraged me in every point of my life.

Abstract

Natural language is ambiguous in different online social media platforms. It contains influential information about different perspectives for customers' feedback. These digitally stored feedbacks are available in abundance and in a myriad of forms in the internet especially social networks platforms e.g. Twitter. It is impossible to manually sift and analyze the needed information. This tremendous growth in data increases the need for reasonable filtering and knowledge mining approaches for decision support systems. For example, organizations need to take appropriate actions and decisions to build their brand and develop their online communities based on customer feedbacks and improve their profits.

Sentiment analysis is non-trivial extraction of implicit user opinions from documents. Emotion detection is a task in sentiment analysis, it deals with the extraction of discrete user attitude from the input data. Analyzing discrete users' attitudes in tweets is a quite difficult task. Researchers face a lot of challenges and obstacles in enhancing accuracy of automatic emotion detection. These challenges can be listed in three main points bipolarities in the same input text, inadequate representative feature vectors, and limitations of adequate annotated lexicons and datasets. Bipolarity challenge is due to natural language ambiguity especially slang and unstructured text, where users rephrase words to express their feelings in various and complex forms. Inadequacy of feature vectors is lack of representative semantic relatedness between input text and target emotion classes. Then, limitation of annotated lexicons is due to the cost of effort and time to build well classified annotated lexicons, which entail widespread acronyms and hashtags.

This research contributes in improving emotion detection accuracy in text. Firstly, it extracts tokens that imply emotions. Secondly, it converts preprocessed tokens to representative feature vectors which reflect the semantic relatedness between input text and the classified emotion using well

annotated lexicons, word embeddings, unigrams, and term frequency techniques. Thirdly, these feature vectors are trained over Naïve Bayes, Support vector machine, Multi-layer perceptron and Convolutional neural networks algorithms to generate five multi-label emotion detection models. Finally, three ensemble techniques are applied to generated models to produce the final implied emotions. This research proposed representative feature vectors and applied ensemble techniques on the implemented classifiers to enhance the previous achieved accuracy by five percentage. The real dataset is used in training and evaluating the proposed system. The proposed system has hamming score 0.59 and average f1 score 0.65.

Published Papers

- 1 Soha S. Ibrahim, Mostafa M. Aref, “NLP in Social Media: An Overview”, the Fifteenth Conference on Language Engineering (ESOLEC’ 2015), December 9-10, 2015, Cairo, Egypt.
- 2 S. S. Ibrahiem, K. A. Bahnasy, M. M. Morsey, M. M. Aref, “Feature Extraction Enhancement in users’ Attitude Detection”, The International Journal of Intelligent Computing and Information Science (IJICIS), Volume 18, Issue 2, Page 1-13, Spring 2018.
- 3 S. S. Ibrahiem, S.S. Ismail, K. A. Bahnasy, M. M. Aref, “Convolutional neural network multi-emotion classifiers”, Jordanian Journal of Computers and Information Technology, 5(2): 97-108, 2019, doi: 10.5455/jjcit.71-1555697775.
- 4 S. S. Ibrahiem, S.S. Ismail, K. A. Bahnasy, M. M. Aref, “Multi-Emotion Classification Evaluation via Twitter”. The 9th IEEE International Conference on Intelligent Computing and Information System (ICICIS), Cairo, Egypt, December 8-19, 2019.
- 5 S. S. Ibrahiem, S.S. Ismail, K. A. Bahnasy, M. M. Aref, “A Case Study in Multi-label Emotion Classification via Twitter”, 12th International Conference on Electrical Engineering (ICEENG), Cairo, Egypt, July 7-9, 2020.

Table of Contents

Chapter One	1
1. Introduction	1
1.1 Problem Definition	3
1.2 Objectives	4
1.3 Contributions	4
1.4 Thesis Organization	5
Chapter Two	6
2. Background & Related Work	6
2.1 Background	6
2.1.1 Sentiment Analysis Applications	6
2.1.2 Opinions Methods Categorization	7
2.1.3 Sentiment Analysis Levels	9
2.1.4 Opinion Types	9
2.1.5 Opinion Structures	10
2.1.6 Sentiment Analysis Challenges	10
2.2 Sentiment Analysis Workflow	13
2.2.1 Data Collection	13
2.2.2 Data Preparation	14
2.2.3 Sentiment Classification	15
2.2.4 Opinion summarization	23
2.3 Emotion Detection	23
2.3.1 Single – label Detection	25
2.3.2 Multi - label Detection	25
2.4 Evaluation Metrics	26
2.4.1 Micro Average F score	27
2.4.2 Hamming Score	27
2.4.3 Hamming Loss	27
2.5 Related Work	28
2.5.1 Enabling Affective Intuition for Concept-Level Sentiment Analysis ..	28
2.5.2 Expanded Aspect and Polarity-Ambiguous Word Lexicon	28

2.5.3	Analyzing domain suitability of a sentiment lexicon	29
2.5.4	Online Paper Review Analysis	29
2.5.5	A Constrained Optimization Approach	30
2.5.6	A Case-Study on Arabic Social Media Posts	30
2.5.7	Sentiment, Emotion, Purpose, and Style in Electoral Tweets	31
2.5.8	Learning word dependencies	31
2.5.9	Detecting Stance in Tweets.....	32
2.5.10	Exploring Fine-Grained Emotion Detection in Tweets	32
2.5.11	Metaphor as a Medium for Emotion	32
2.5.12	Deep learning-based document modeling.....	33
2.5.13	Fine-Grained Emotion Analysis.....	33
2.5.14	Emotion Detection and Analysis on Social Media.....	34
2.5.15	Emotion Detection from Speech and Text.....	34
2.5.16	A Corpus for Sentiment Analysis and Emotion Recognition for a Learning Environment	35
2.5.17	Two Step Approach for Emotion Detection on Twitter Data.....	35
2.6	Evaluation.....	35
2.7	Summary	42
Chapter Three		43
3.	Users' attitude System Architecture	43
3.1	The Emotion Detection Workflow.....	43
3.2	Data Preparation Phase.....	44
3.2.1	Data Preprocessing:.....	45
3.2.2	Feature extraction.....	48
3.3	Emotion Detection	51
3.3.1	Multi Modal Emotion Detection system.....	51
3.3.2	Ensemble Techniques	61
3.4	Case study I	61
3.5	Case Study II.....	64
3.6	Summary	66
Chapter Four		67

4. Implementation, Results, & Evaluation	67
4.1 Datasets & Lexicons	68
4.1.1 Annotated Datasets.....	68
4.1.2 Emotion Lexicons	68
4.1.3 Tools.....	69
4.2 Preprocessing & Feature extraction	70
4.2.1 Frequency distribution.....	70
4.2.2 Word-to-Vector Similarity.....	72
4.2.3 Average probability lexicons distribution.....	75
4.3 Four Emotion Detection Classifiers	76
4.4 Eight Emotion Detection Classifiers	85
4.5 Ensemble Techniques on Four Emotion Detection Classifiers.....	95
4.6 Eight Emotion Detection Classifiers Evaluation.....	97
4.7 Discussion.....	98
4.8 Summary	100
Chapter Five	101
5. Conclusion & Future Work	101
5.1 Conclusion	102
5.2 Future Work.....	103
References	104

List of Tables

Table 2-1: Opinion Dimensions	8
Table 2-2: Comparison of SA approaches	22
Table 2-3: Emotions Common lists in ED Methods.....	24
Table 2-4: Supervised Researches in SA since 2015.....	36
Table 2-5: Supervised Researches in SA since 2015 (Continued)	36
Table 2-6: Lexicon-based Researches in SA since 2015	37
Table 2-7: Hybrid Researches in SA since 2015.....	38
Table 3-1: Preprocessed Tokens for The Input Tweet.....	62
Table 3-2: Feature Values for The Input Tweet	63
Table 3-3: VAD Feature values for The Input Tweet.....	63
Table 3-4: Weka Models Predictions of The Input Tweet	63
Table 3-5: Sample Tweets from SemEval2018.....	64
Table 3-6: First Tweet feature vector	65
Table 3-7: Users' Attitudes Predicted in Sample Tweets	65
Table 4-1 : Naïve Bayes Classification Report in Four Emotion Detection Classifier	77
Table 4-2 : SVM Classification Report in Four Emotion Detection Classifier	79
Table 4-3 : KNN Classification Report in Four Emotion Detection Classifier	80
Table 4-4 : MLP Classification Report in Four Emotion Detection Classifier	81
Table 4-5 : Parallel CNN Classification Report in Four Emotion Detection Classifier	83
Table 4-6 : Single CNN Classification Report in Four Emotion Detection Classifier	85
Table 4-7 : Gaussian NB Classification Report in Eight Emotion Detection Classifier	87
Table 4-8 : SVM Classification Report in Eight Emotion Detection Classifier.....	88
Table 4-9 : KNN Classification Report in Eight Emotion Detection Classifier.....	90
Table 4-10 : MLP Classification Report in Eight Emotion Detection Classifier	91
Table 4-11 : Parallel CNN Classification Report in Eight Emotion Detection Classifier	93
Table 4-12 : Single CNN Classification Report in Eight Emotion Detection Classifier	95
Table 4-13 : Four Emotion Detection Classifier Evaluation.....	96
Table 4-14: Eight Emotion Detection Classifiers Evaluation	97

List of Figures

Figure 1.1: Google Trends of search strings “sentiment analysis” and “customer feedback”	2
Figure 2.1: The Accuracy Improvement in SA Challenges	12
Figure 2.2: General SA Workflow	13
Figure 2.3: SA Classification Approaches	15
Figure 3.1 Emotion Detection Workflow	44
Figure 3.2: Preprocessing Data Flow Diagram	47
Figure 3.3: Training Workflow in Multi Emotion Detection	52
Figure 3.4: Testing Workflow in Multi Emotion Detection	53
Figure 3.5: Architecture of First Proposed CNN Multi – Emotion Detection	56
Figure 3.6: Architecture of Second Proposed CNN Multi – Emotion Detection	58
Figure 3.7: Pseudocode of Emotion Detection System	60
Figure 4.1: Explanatory Analysis of Term Frequency for each emotion class: (a) angry tokens, (b) anticipation tokens, (c) disgust tokens, (d) fear tokens, (e) joy tokens, (f) sad token, (g) surprise tokens, & (h) trust tokens	71
Figure 4.2: Explanatory Analysis between two emotion classes word embeddings: (a) Sad and Fear, and (b) anger and disgust	73
Figure 4.3: Explanatory Analysis between two emotion classes word embeddings: (a) Joy and trust, and (b) surprise and anticipation.....	74
Figure 4.4: Gaussian NB Four Emotion Detection Learning Curve	76
Figure 4.5: SVM Four Emotion Detection Learning Curve	78
Figure 4.6: KNN Four Emotion Detection Learning Curve	79
Figure 4.7: MLP Four Emotion Detection Learning Curve	80
Figure 4.8: Parallel CNN Four Emotion Detection Learning Curve	82
Figure 4.9: Single CNN Four Emotion Detection Learning Curve	84
Figure 4.10: Gaussian NB Eight Emotion Detection Learning Curve.....	86
Figure 4.11: SVM Eight Emotion Detection Learning Curve.....	87
Figure 4.12: KNN Eight Emotion Detection Learning Curve.....	89
Figure 4.13: MLP Eight Emotion Detection Learning Curve	90
Figure 4.14: Parallel CNN Eight Emotion Detection Learning Curve	92
Figure 4.15: Single CNN Eight Emotion Detection Learning Curve	94

List of Abbreviations

BOW – Bags of Words

CNN – Convolutional Neural Network

ED – Emotion Detection

KNN – K Nearest Neighbor

ME – Maximum Entropy

MLP – Multi Layer Perceptron

NB – Naïve Bayes

NE – Named Entity

NLP – Natural Language Processing

NLTK – Natural Language Toolkit

NN – Neural Network

NRC – National Research Council (in Canada)

POS – Part of Speech

ReLU – Rectified Linear Units

SA – Sentiment Analysis

SemEval – International Workshop on Semantic Evaluation

SO – Semantic Orientation

SVM – Support Vector Machine

VAD – Valence Arousal, Dominance

W2V – Word to Vector

Chapter One

1. Introduction

Opinions are central to all human activities and are key influencers of their behaviors. Humans' beliefs and perceptions of reality, and choices they make, are to a considerable degree, conditioned upon how they see and evaluate the world. In recent years there is a huge increase of the role of social networks in the formation of both the content and the audience of modern mass media. Customer Feedback from social media happens when customers make comments on Twitter, Facebook, and a number of other social media, blogs, and review sites. This feedback is either available to everyone, or may be exclusive to the customer's followers. Typically, business organizations have to learn their customers' feedback. They need to scan hashtags and keywords to find this feedback, so that they can take actions to turn negative experience into positive outcome. The data gathering, sifting, and analyzing processes on customers feedback are impossible to be manually manipulated, it should be automated as far as possible to limit cost of effort and time.

Opinions and its related concepts such as feedbacks, sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining [1]. Sentiment analysis (SA) is the computational process that study people's opinions, attitudes and emotions towards an entity, this entity can be individuals, events or topics [2]. There are also many names and slightly different tasks, e.g.: sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. Sentiment analysis is a laborious task for performing with computers and algorithms. It is a Natural Language Processing (NLP) problem. It touches every aspect in NLP e.g. co-reference