بسم الله الرحمن الرحيم

MONA MAGHRABY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

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 Ibrahiem 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 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

`

`

`

List of Tables

`

List of Figures

`

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