بسم الله الرحمن الرحيم

MONA MAGHRABY

شبكة المعلومات الجامعية

التوثيق الالكتروني 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 تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

MONA MAGHRABY

AIN SHAMS UNIVERSITY
Faculty of Computer and Information Sciences
Information Systems Department

Feature-based Approach for Sentiment Analysis of Social Networks

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of M.Sc. in Computer and Information Sciences
To
Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University

By

Nagwa Moustafa Kamal Saeed
Teaching Assistant at Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Under the Supervision of

Prof. Dr. Tarek Fouad Gharib
Head of Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Prof. Dr. Nagwa Lotfy Badr
Information Systems Department
Dean of the Faculty of Computer and Information Sciences
Ain Shams University

Dr. Nivin Atef Helal
Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

2020

Acknowledgment

First and foremost, I would like to thank God Almighty for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily.

I wish to express a deep sense of gratitude and appreciation to Prof. Dr. Tarek Gharib, Prof. Dr. Nagwa Badr and Dr. Nivin Atef for their supervision, continuous support, encouragement, invaluable guidance and skillful suggestions which deeply contributed to the completion of this research study.

Finally, I would like to express my deep sense of gratitude towards my beloved family who are the biggest source of my strength and of course my prime source of ideas. They have all made a tremendous contribution in helping me reach this stage in my life. Had it not been for their unflinching insistence and their support to me, my dream of getting this degree would have remained a mere dream. Therefore, I would like to thank all of them for their endless love, continual support, prayers, patience, understanding and encouragement throughout these years and for their believing in me that I can finish my research study in time.

Abstract

In the last few years, online reviews where individuals express their thoughts, interests, experiences and opinions have majorly spread over the internet. Sentiment analysis field of study has evolved to analyze these online reviews and provide valuable insights for both individuals and organizations that may help them in making decisions. Unfortunately, the performance of sentiment analysis process is affected by the nature of online reviews' content that may contain emoticons and negation words. Moreover, spam reviews have been written for the purpose of deceiving others. These spam reviews may greatly influence online marketing and prevent both individuals and organizations from concluding real ideas about certain services or products. Therefore, there is a need to develop an approach that considers these issues.

In this thesis, an enhanced approach for sentiment analysis is proposed which aims to enhance the performance of classifying reviews based on their features and assigning an accurate sentiment score to each feature. This enhanced approach is achieved by handling negation, detecting emoticons, and detecting spam reviews using a combination of different types of properties which leads to achieving better predictive performance. Moreover, this approach examines the impact of using three different feature extraction methods on the performance of sentiment classification which are extracting all nouns, extracting only the nouns that occur frequently, and extracting frequent nouns by applying Apriori algorithm.

Several experiments have been carried out to validate the effectiveness of the proposed approach. The performance of the proposed approach has been measured using different types of evaluation metrics which are accuracy, precision, recall, and f1 score. The proposed approach has been verified against three datasets of different sizes. The experimental results showed the efficiency of the proposed approach in detecting spam reviews, classifying reviews based on their features and assigning an accurate sentiment score to each feature. The proposed approach achieves a maximum accuracy of about 99.06% in detecting spam reviews and outperforms the existing related works with an average value of 13.35% for accuracy. The proposed approach achieves as well a maximum accuracy of about 97.13% in classifying reviews after considering the three main challenges: negation handling, emoticons detection, and spam reviews detection together and after employing “extracting frequent nouns by applying Apriori algorithm” as a feature extraction method, where there is an improvement in accuracy value of about 29.72%, and a great saving in the feature space by 96.9% versus when not considering these three main challenges together along with this feature extraction method.

List of Publications

1. Saeed NMK, Helal NA, Badr NL, Gharib TF (2018) The Impact of Spam Reviews on Feature-based Sentiment Analysis. In: Proc. - 2018 13th Int. Conf. Comput. Eng. Syst. ICCES 2018. IEEE, pp 633–639.

2. Saeed NMK, Helal NA, Badr NL, Gharib TF (2020) An enhanced feature-based sentiment analysis approach. Wiley Interdiscip Rev Data Min Knowl Discov 10:1–20.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CNN	Convolutional Neural Network
DOSC	Deceptive Opinion Spam Corpus
DT	Decision Tree
Emoticon	Emotion Icon
FbSA	Feature-based Sentiment Analysis
IDE	Integrated Development Environment
KNN	K-Nearest Neighbors
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
NLP	Natural Language Processing
NN	Neural Networks
NYC	New York City
POS	Part of Speech
RF	Random Forest
SVM	Support Vector Machine
ZIP	Zone Improvement Plan

Chapter 1
Introduction

Nowadays, there is a major increase in the number of people using the internet mainly social networking sites. Social networking sites provide a virtual community which allows people to interact with each other and share their thoughts, interests, experiences and opinions about different topics. This situation has led to the formation of a very huge amount of online reviews which can be exploited in various fields such as healthcare, politics, marketing, sociology, entertainment, etc. These online reviews have become a valuable source of information that people can refer to in order to judge the quality of a certain service or product while making decisions to use or purchase this service or product. Manually processing or analyzing these online reviews would be a very difficult task as it would be time consuming and may lead to inaccurate decisions too. This naturally led to the emergence of the field of study which is known as opinion mining and sentiment analysis [1].

Opinion mining and sentiment analysis are considered as subfields of natural language processing (NLP), information retrieval and text mining. Opinion mining is the process of extracting users’ opinions and thoughts expressed on entities or features/aspects of entities from unstructured texts, while sentiment analysis is the process of analyzing the opinionated text and determining its polarity in an automated manner [2].

1.1. Motivation

In these days, the need for automatically tackling and analyzing large volumes of individuals' online reviews is growing day after day. The analysis of these online reviews is performed in order to infer useful information that can be exploited to help both individuals and organizations in concluding real ideas and gaining experiences about certain services or products. Moreover, this information can be helpful in making a decision concerning buying a product or in enhancing the quality of services or products. Unfortunately, the process of automatically analyzing and detecting the sentiment in these online reviews faces several challenges such as (spam reviews detection, emoticons detection, negation handling, etc.) which all affect its overall performance and lead to low accuracy in the classification of the users' sentiments.

Furthermore, previous different studies either have applied sentiment analysis on different structural levels (Document level, Sentence level and Feature level) without considering any of the challenges mentioned above or have applied sentiment analysis with considering only one challenge of the above mentioned ones or detected spam reviews only without considering their influence on the performance of sentiment classification. So, these studies generally handled some of the above mentioned challenges but not all of them together which led to low accuracy in the classification of the users' sentiments. Therefore, there is a need to develop a sentiment analysis approach that considers these challenges.