



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 تغيرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



MONA MAGHRABY



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

Abstract	I
List of Publications.....	II
Table of Contents	III
List of Figures.....	V
List of Tables	VII
List of Abbreviations.....	VIII
Chapter 1: Introduction.....	1
1.1 Motivation	2
1.2 Objective	3
1.3 Thesis Organization	3
Chapter 2: Opinion Mining and Sentiment Analysis Background.....	5
2.1 Opinion Mining and Sentiment Analysis	5
2.2 Structural Levels of Sentiment Analysis	5
2.3 Opinion Mining and Sentiment Analysis Challenges	7
2.4 Summary	9
Chapter 3: Related Work	10
3.1 Studies Performing Sentiment Analysis	10
3.2 Studies Detecting Spam Reviews Only	15
3.3 Summary	17
Chapter 4: Enhanced Feature-based Sentiment Analysis Approach	18
4.1 Preprocessing.....	22
4.2 Features and Sentiment Terms Extraction	22
4.2.1 Features Extraction	23

4.2.2	Opinion Words Extraction	23
4.2.3	Emoticons Extraction.....	24
4.3	Features' Sentiment Scores Calculation	24
4.3.1	Opinion Words' Sentiment Scores Calculation	26
4.3.2	Negation Handling	27
4.3.3	Emoticons' Sentiment Scores Calculation.....	27
4.4	Spam Reviews Detection	29
4.4.1	Property Processing	29
4.4.2	Spam/Truthful Reviews Classification	32
4.5	Polarity Classification	37
4.6	Summary	38
Chapter 5: Evaluation, Experimental Results and Discussions		39
5.1	Dataset.....	39
5.2	Evaluation Metrics.....	40
5.3	Experimental Results	42
5.3.1	Features Extraction	42
5.3.2	Negation Handling and Emoticons Detection.....	44
5.3.3	Spam Reviews Detection	48
5.3.4	Enhanced Feature-based Sentiment Analysis	61
5.4	Discussions.....	65
5.5	Summary	70
Chapter 6: Conclusion and Future Work.....		71
6.1	Conclusion.....	71
6.2	Future Work	72
References.....		73
Arabic Summary		78

List of Figures

Figure 4.1: Graphical illustration for the proposed feature-based sentiment analysis approach	19
Figure 4.2: Overview of the proposed approach	20
Figure 4.3: Enhanced feature-based sentiment analysis algorithm.....	21
Figure 4.4: Features’ sentiment scores calculation algorithm	25
Figure 4.5: Overview of the different types of properties used for spam reviews detection	31
Figure 4.6: Spam reviews detection using rule-based classification algorithm	33
Figure 4.7: Spam reviews detection using machine learning classification algorithm.....	36
Figure 4.8: Polarity classification algorithm.....	37
Figure 5.1: Confusion matrix	40
Figure 5.2 (a): Performance of FbSA with and without considering (negation handling + emoticons detection) on “DOSC” dataset	45
Figure 5.2 (b): Performance of FbSA with and without considering (negation handling + emoticons detection) on “YelpNYC” dataset.....	45
Figure 5.2 (c): Performance of FbSA with and without considering (negation handling + emoticons detection) on “YelpZIP” dataset	46
Figure 5.3: The impact of using different set of properties on the performance of the rule-based spam reviews detection method.....	49
Figure 5.4 (a): Accuracy of the machine learning classification spam reviews detection method on “DOSC” dataset	52
Figure 5.4 (b): Accuracy of the machine learning classification spam reviews detection method on “YelpNYC” dataset	53

Figure 5.4 (c): Accuracy of the machine learning classification spam reviews detection method on “YelpZIP” dataset	53
Figure 5.5 (a): Performance of spam reviews detection methods on “DOSC” dataset	58
Figure 5.5 (b): Performance of spam reviews detection methods on “YelpNYC” dataset	58
Figure 5.5 (c): Performance of spam reviews detection methods on “YelpZIP” dataset	59
Figure 5.6 (a): Performance of FbSA (negation handling + emoticons detection) before and after spam reviews detection on “DOSC” dataset	62
Figure 5.6 (b): Performance of FbSA (negation handling + emoticons detection) before and after spam reviews detection on “YelpNYC” dataset	62
Figure 5.6 (c): Performance of FbSA (negation handling + emoticons detection) before and after spam reviews detection on “YelpZIP” dataset	63

List of Tables

Table 4.1: Example of opinion words and their sentiment scores	26
Table 4.2: Different purposes of using emoticons and their influence on enhancing the sentiment classification process	28
Table 5.1: The performance of feature-based sentiment analysis (FbSA) on using different feature extraction methods	44
Table 5.2: The impact of considering negation handling and emoticons detection on the performance of feature-based sentiment analysis (FbSA)	47
Table 5.3: The impact of using different set of properties on the performance of the rule-based spam reviews detection method.....	50
Table 5.4: The performance of spam reviews detection on using different machine learning techniques and properties' set.....	55
Table 5.5: Features' sentiment scores before and after detecting spam reviews ...	57
Table 5.6: Comparison between state-of-the-art spam reviews detection results and our proposed method.....	61
Table 5.7: The impact of spam reviews detection on the performance of FbSA while considering negation handling and emoticons detection	64

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.