شبكة المعلومات الجامعية

بسم اللّه الرحمن الرحيم

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم باللّه 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

شبكة المعلومات الجامعية

بعض الوثائق الأصلية تالفة

بالرسالة صفحات
لم ترد بالأصل

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

A Sentiment Analysis Approach for Arabic Texts

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

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

Under the Supervision of

Tarek Fouad Gharib
Professor, Head of Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Sherine Rady Abdel Ghany
Associate Professor, Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

2020

Acknowledgment

Foremost, I thank God for providing me with the patience and guidance to finish this work.

Many thanks and sincere gratefulness to Prof. Dr. Tarek Fouad Gharib and Dr. Sherine Rady Abdel Ghany for their supervision, continuous encouragement, rich discussion, precious comments, and suggestions throughout the research and the thesis work.

Last but not the least important, I would really like to explicit my deep gratitude towards my beloved family who are the most important source of my power and of course my prime source of thoughts. They have all made a massive contribution in availing me reach this stage in my life. Had it not been for their unflinching insistence and their assist to me, my dream of getting this degree might have remained just a dream. Therefore, I would relish the chance to thank all of them for their limitless love, prayers, continual assist, patience, understanding, and encouragement throughout those years and for their believing in me that I am able to end my research in time.

Abstract

Nowadays, individuals express their experiences and opinions through online reviews. These reviews influence online marketing and provide a guide for potential customers allowing them to reach real knowledge about products/services while making decisions. Sentiment analysis is the process of analyzing opinions expressed in textual reviews automatically. The efficiency of this process is affected by the spammed opinion information, and by the set of representative features extracted from the reviews. Prior spam detection researches and most sentiment classification studies integrating dimensionality reduction have focused on English texts, with less attention to other languages, including Arabic. Huge amounts of Arabic data have been generated due to the huge population of Arab world; and despite that, the aforementioned technical gaps still exist for such language.

In this thesis, a supervised learning approach for Arabic reviews' sentiment classification is proposed. This approach utilizes optimal compact features that depend on a well representative feature set coupled with feature reduction technique, which provides high accuracy and time/space savings. The employed feature set includes a triple combination of N-gram features and positive/negative N-grams counts features obtained after negation handling. Two different linear transformation methods are studied; Principal Component Analysis (PCA) as an unsupervised method and Latent Dirichlet Allocation (LDA) as a supervised method. Spam detection is also employed as a prior process to the classification to increase its robustness. Four different Arabic spam reviews detection methods are proposed while putting more focus towards the construction and evaluation of ensemble approaches, which integrate rule-based classification and machine learning techniques, and with the use of content-based features that depend on N-gram features and negation handling.

The proposed Arabic sentiment classification approach and Arabic spam reviews detection methods have been assessed by conducting several experiments. The sentiment classification approach has been evaluated on five Arabic opinion text datasets, of different domains and with varying sizes (1.6K up to 94K reviews). The approach has been experimented for classifying sentiments in two (positive/negative) and three (positive/negative/ neutral) class problems. Accuracy values for the feature reduction-based sentiment analysis approach occurred in the range 95.5−99.8% for 2-class problem and 92−97.3% for 3-class problem and outperformed existing related works by far of 23% for accuracy. The LDA feature reduction outperformed PCA by an average of 4.34% in accuracy. The results also demonstrated significant improvement with 24% increase in accuracy, 93% savings in the feature space, and 97% decrease in the classification execution time. The four spam reviews detection methods have been evaluated on two Arabic opinion text datasets of different sizes (1.6K and 94K reviews). The results indicated the efficiency of the ensemble method, where it achieved accuracy values of 95.25% and 99.98% for the two experimented datasets and outperformed existing related works by far of 25% for accuracy.

List of Publications

1. R. M. K. Saeed, S. Rady, and T. F. Gharib, “An Ensemble Approach for Spam Detection in Arabic Opinion Texts,” J. King Saud Univ. - Comput. Inf. Sci. 2019. https://doi.org/10.1016/j.jksuci.2019.10.002.

2. R. M. K. Saeed, S. Rady, and T. F. Gharib, “Optimizing Sentiment Classification for Arabic Opinion Texts”. Cognit. Comput. 2020. https://doi.org/ 10.1007/s12559-020-09771-z.

Table of Contents

Abstract I

List of Publications II

Table of Contents III

List of Figures VI

List of Tables IX

List of Abbreviations XI

Chapter 1: Introduction 1

1.1 Motivation 3

1.2 Objective 4

1.3 Thesis Contributions 4

1.4 Thesis Organization 5

Chapter 2: Background 7

2.1 Introduction to Sentiment Analysis 7

2.2 Sentiment Analysis Techniques 8

2.2.1 Lexicon-based Technique 8

2.2.2 Machine Learning Technique 9

2.2.3 Hybrid Technique 12

2.3 Arabic Language Sentiment Analysis 12

2.4 Summary 13

Chapter 3: Related Work 14

3.1 Studies Detecting Spam Reviews 14

3.2 Studies Performing Sentiment Classification 17

3.3 Summary 22

Chapter 4: A content-based Approach for Arabic Spam Reviews Detection 23

4.1 Pre-processing 25

4.2 Feature Extraction 26
4.2.1 N-gram Feature Extraction 26
4.2.2 Negation Handling 27
4.2.3 Content-based Feature Extraction 27
4.3 Spam Reviews Detection 28
4.3.1 Rule-based Classifier 28
4.3.2 Machine Learning Classifiers 29
4.3.3 Majority Voting Ensemble Classifier 29
4.3.4 Stacking Ensemble Classifier 30
4.4 Experimental Results and Discussion 31
4.4.1 Datasets 31
4.4.2 Evaluation Metrics 32
4.4.3 Experimental Results 33
4.4.4 Analysis and Discussion 44
4.4.5 Comparison with State-of-the-Art Spam Reviews Detection Approaches 46
4.5 Summary 48
Chapter 5: Optimized Feature Generation for Arabic Sentiment Classification 49
5.1 Pre-processing 50
5.2 Spam Reviews Detection 53
5.3 Optimized Features Generation 53
5.3.1 Feature Extraction 53
5.3.2 Feature Selection 56
5.4 Sentiment Classification 58
5.5 Experimental Results and Discussion 59
5.5.1 Datasets 59
5.5.2 Evaluation Metrics 60

5.5.3 Experimental Results 61
5.5.4 Analysis and Discussion 76
5.5.5 Comparison with State-of-the-Art Arabic Sentiment Classification Approaches ..... .78
5.6 Summary 81
Chapter 6: Conclusion and Future Work 82
6.1 Conclusion 82
6.2 Future Work 83
References 84
Arabic Summary 91

List of Figures

Figure 4.1: Overview of the proposed work 23

Figure 4.2: Overview of the proposed spam detection approach 24

Figure 4.3: Overview of the majority voting ensemble classifier 30

Figure 4.4: Overview of the stacking ensemble classifier 31

Figure 4.5: Performance of rule-based classifier for Arabic spam detection while using different combinations of N-grams on DOSC 35

Figure 4.6: Performance of rule-based classifier for Arabic spam detection while using different combinations of N-grams on HARD 35

Figure 4.7: Performance of rule-based classifier for Arabic spam detection before and after applying negation handling on DOSC 37

Figure 4.8: Performance of rule-based classifier for Arabic spam detection before and after applying negation handling on HARD 37

Figure 4.9: Accuracy of different machine learning classifiers for Arabic spam detection before and after negation handling on DOSC 41

Figure 4.10: Accuracy of different machine learning classifiers for Arabic spam detection before and after negation handling on HARD 41

Figure 4.11: Performance of majority voting ensemble classifier for Arabic spam detection while working on DOSC and HARD 42

Figure 4.12: Comparing the accuracy of the four spam review detection methods for DOSC and HARD 45

Figure 4.13: Accuracy of stacking ensemble classifier versus varying data sizes 46

Figure 4.14: Comparison between the stacking ensemble approach and some related works on DOSC 47

Figure 5.1: Overview of the proposed sentiment classification approach 50

Figure 5.2 (a): Accuracy of the best machine learning classifier while using different combinations of N-grams when applied to 2-class problem 63

Figure 5.2 (b): F1 Score of the best machine learning classifier while using different combinations of N-grams when applied to 2-class problem 63

Figure 5.3 (a): Accuracy of the best machine learning classifier while using different combinations of N-grams when applied to 3-class problem 64

Figure 5.3 (b): F1 Score of the best machine learning classifier while using different combinations of N-grams when applied to 3-class problem 65

Figure 5.4 (a): Accuracy of the best machine learning classifier before and after negation handling when applied to 2-class problem 67

Figure 5.4 (b): F1 Score of the best machine learning classifier before and after negation handling when applied to 2-class problem 67

Figure 5.5 (a): Accuracy of the best machine learning classifier before and after negation handling when applied to 3-class problem 68

Figure 5.5 (b): F1 Score of the best machine learning classifier before and after negation handling when applied to 3-class problem 69

Figure 5.6: Word cloud for the most related word in positive and negative reviews 71

Figure 5.7 (a): Accuracy of the best machine learning classifier before and after employing feature selection along with the positive/negative N-grams counts features when applied to 2-class problem 73

Figure 5.7 (b): F1 Score of the best machine learning classifier before and after employing feature selection along with the positive/negative N-grams counts features when applied to 2-class problem 73

Figure 5.8 (a): Accuracy of the best machine learning classifier before and after employing feature selection along with the positive/negative N-grams counts features when applied to 3-class problem 74

Figure 5.8 (b): F1 Score of the best machine learning classifier before and after employing feature selection along with the positive/negative N-grams counts features when applied to 3-class problem 75

Figure 5.9: Performance summary for the proposed Arabic sentiment classification approach on the five experimented datasets..... 77

Figure 5.10: Accuracy of the proposed Arabic sentiment classification approach versus varying data sizes..... 78

Figure 5.11 (a): Comparison between the performance of state-of-the-art Arabic sentiment classification and the performance of our proposed work in terms of accuracy..... 80

Figure 5.11 (b): Comparison between the performance of state-of-the-art Arabic sentiment classification and the performance of our proposed work in terms of F1 Score..... 80