بسم الله الرحمن الرحيم

∞∞∞

تم رفع هذه الرسالة بواسطة / هناء محمد علي

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى مسئولية عن محتوى هذه الرسالة.

ملاحظات: ....

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

Developing An Approach For Opinion Mining On Online Reviews

Thesis submitted as a partial fulfilment of the requirements for the degree of Master of Science in Computer and Information Sciences

By

Eman Mahmoud Aboelela Mohamed

Teaching Assistant at Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University.

Under Supervision of

Prof. Rasha Ismail

Professor, Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University.

Prof. Walaa Khaled Ibn El-walid

Professor, Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University.

2022

Contents

List of Figures

List of Tables

Abstract

THROUGH the state-of-the-art digitalization, we can see a massive growth in user-generated content on the web that provides feedback from people on a variety of topics. However, manually managing large-scale user feedback would be a hard task and a waste of time. Therefore, the concept of opinion mining or sentiment analysis is emerged. Opinion mining is a computerized study of individuals' feelings and opinions about an entity or product. Opinion mining is difficult because the users' opinions or reviews are mostly unstructured text, lower quality, noisy, and spam. Thus, there are several challenges facing the opinion mining and causing poor classification performance. Some of these challenges are multiple languages of user reviews, fake reviews, manipulation of emoticons, implicit aspects, spam aspects, and negative words that change the class label of opinion words. In this thesis a semantic-based aspect level opinion mining (SALOM) model is proposed. In order to address some of the previously mentioned challenges. Whereas, SALOM involves the semantic similarity measure and Wordnet ontology to extract the product related domain aspects. Moreover, it considers other types of product aspects such as aspect-related synonym, hyponym, and hypernym. Not only that, but the proposed model also considers the reviews that contain product aspects only. In addition, SALOM model considers negation words that affect the performance of the opinion mining process. Five different datasets are used to evaluate the proposed SALOM model. SALOM achieved promising results compared to other methods. Where, the performance reached 91.6% for accuracy, 93% for recall, 95.8% for precision, and 93.3% for f-measure.

Acknowledgements

I want to thank all the people who helped me to complete this work.

Firstly, I would like to express my deepest thankfulness to my supervisors: Prof. Rasha Ismail and Prof. Walaa Khaled. I am very thankful for their advice, guidance, support, and motivation. They spent much time revising my work. I wish to send my sincere regards and appreciation to my family for being my back. My parents, brothers, husband's mother. Special thanks to my small family: husband and children for being the source of support and motivation.

Above all, I would like to thank God because He empowered me to finish this work successfully.

Chapter 1

Introduction

1.1 Overview

People all over the world use social networking sites such as Facebook, Twitter, Instagram, etc. especially during the COVID-19 pandemic. In order to share their experiences, opinions and thoughts on specific topics with others. These opinions and thoughts are an important reference to customers in evaluating the products and taking decisions on which product to purchase. Moreover, the vendors use these opinions to know the positive and negative aspects of their products. Therefore, opinions collection and analysis play a vital role in the product design and marketing. Opinion mining is used to monitor the peoples' feelings and emotions toward specific products or services Arunkarthi and Gandhi [2]. Opinion mining is useful for filtering the users' opinions and generating a quick summary from a large amount data in an acceptable time. However, the large number of opinions is considered as beneficial as challenging to analyze. There are different kinds of opinion analysis which are manual analysis, keyword analysis, and natural language analysis Usharani [59]. Manual analysis is time consuming and may fail or cause incorrect decisions as it depends on human observation and interpretation. Keyword analysis classifies the individual terms into positive or negative and measures the overall score of the comment. Natural language analysis analyzes the language meaning and is also called text analytic and computational linguistics. Opinion mining is used in dif-

ferent fields of life, such as e-learning, commerce, politics, finance, purchasing items, entertainment and others. Opinion mining is made up of three main components: opinion holder, opinion object, opinion polarity Neshan and Reza [34], Sonia [54]. The opinion holder is the opinion owner and may be an individual or organization. Opinion object is the property or aspect that can be expressed by the opinion holder. Opinion polarity is a class label for positive or negative opinion. There are two main types of web-based reviews which are structured and unstructured Khan et al. [20]. Structured reviews are written in the form of pros and cons to determine what is positive and negative in the service or product. Unstructured reviews are written in a human natural language so, they may contain untruthful words. The textual information has two classes Mishra and Jha [30]: facts and opinions. Facts are objective expressions about events, entities and properties. Opinions are subjective expressions that reflect the people's opinions, emotions and sentiments towards entities and their properties.

Opinion mining uses information retrieval and natural language processing techniques to analyse opinions. Therefore, the opinion mining analyses the reviews according to 3 Levels of analysis, namely document/review level, sentence/phrase level and aspect/feature levelSharma et al. [46], Ligthart et al. [21]. The document level opinion mining, which classifies the whole document or review with positive or negative based on the number of negative and positive sentiment words that appeared in the review. At this level, the review is assumed to have opinions related to a single object. Although, the same review may contain different objects with different opinions of the same person. Therefore, this level of analysis is not accurate for decision making. The opinion mining at sentence level considers each sentence as a separate source that includes a single opinion. It classifies each sentence as either positive or negative. The product aspects are defined as the attributes or properties of the product. The aspects are categorized into implicit and explicit aspects. Implicit aspects are not explicitly written in the review, but derive from the meaning. Explicit aspects are explicitly located in the review. Thus, in aspect level opinion mining, the review is classified into positive and negative labels according to the product aspects appearing in the review

sentences. Whereas the same review may have different aspects with different sentiments. The aspect level opinion mining is the most valuable level of analysis in decision-making. Since it determines the points of strength and weakness of a product or service.

The aspect level opinion mining has two main steps. The first one is the aspect extraction step where the explicit aspects of the product are extracted from the unstructured users' reviews or comments. The second step is the review classification based on the extracted product aspects. Consider the following review from the Nokia 6610 dataset, "the fm radio is cool.". Firstly, "fm radio" is extracted as an aspect of mobile and ("cool") as opinion word. Then, the review is classified according to "fm radio" as positive because "cool" is a positive oriented sentiment.

There are two opinion mining techniques used for the review classification step, namely the lexicon-based technique and the machine learning technique Kamble and Itkikar [18], Onan [37]. Figure 1.1 presents a tree for the opinion mining techniques.

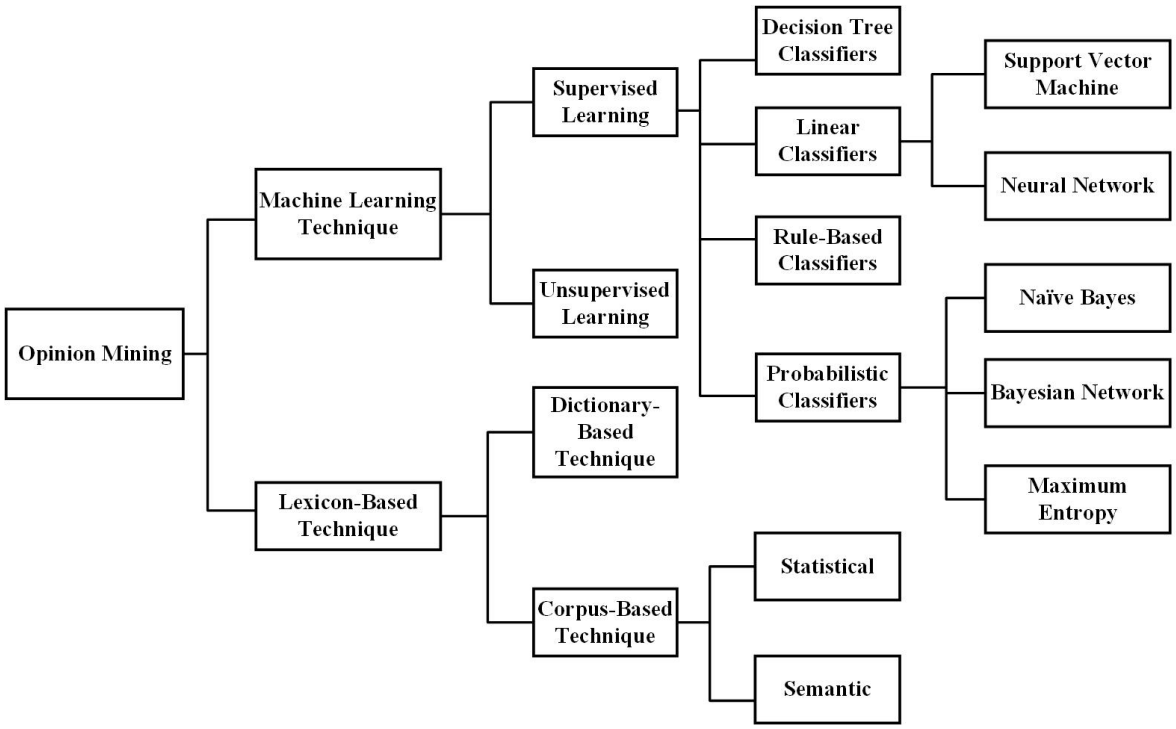

Figure 1.1: A tree of review classification techniques.

As shown in Figure 1.1, the machine learning technique has two categories which are supervised and unsupervised learning. The supervised learning has a labeled trained dataset and test dataset. However, the

unsupervised learning does not have a training dataset so, it uses unlabeled data and finds its hidden patterns. Moreover, the machine learning has a modern research trend with deep learning where high performance classification models are built using multiple layers of nonlinear data representations and supervised/unsupervised learning features of data Onan [37]. In addition, two techniques are shown in Figure 1.1 under the lexicon-based approach which are the dictionary-based technique and the corpus-based technique. In the dictionary-based technique, a dictionary of words is built manually; each word is assigned a positive or a negative value. Afterwards, the dictionary is extended by the synonyms and antonyms of words that are extracted from Wordnet ontology. This extension is terminated when there are no new words to be added to the dictionary Priyavrat and Singh [40]. But, in the corpus-based technique, a sentiment lexicon such as SentiWordnet is used along with a labeled list of words.

1.2 Problem Definition

The performance of the aspect-based opinion mining process through the unstructured reviews is affected by the following:

- The existence of negation words which causes misclassification of opinion words. Whereas, a negation word is able to change the polarity of all its surrounding words in the sentence from positive to negative and vice versa.
- Spam aspects which are less important and not relevant to the product. These aspects affect the performance of review classification.
- Fake reviews which known as spam tendentious reviews documented by anonymous one. These reviews aim to promote or downgrade specific products or services for some business needs.

So, There is a need to build a model that addresses these issues.

1.3 Objective

This work aims to develop a lexicon-based technique to build a model which improves the efficiency of the aspect-based review classification by handling negation words, spam aspects and fake reviews.

1.4 Contribution

In this thesis, We build a model, namely the Semantic-based Aspect Level Opinion Mining SALOM model. The proposed model uses the semantic similarity and Wordnet Ontology to find the domain product aspects and avoid spamming. SALOM considers only the reviews that contain a product aspect to avoid the fake reviews. Additionally, it considers different types of product aspects such as aspects-related synonyms, hypernyms, and hyponyms to process more reviews. These aspects' types are extracted based on Wordnet ontology Sharma et al. [46], and semantic similarity Wan and Angryk [60]. Moreover, negations are considered in the proposed model to accurately classify the reviews. To obtain more precise polarity scores, the proposed SALOM uses a labeled sentiment corpus of the most known positive and negative sentiments along with a sentiment lexicon. Not only that, but also SALOM can deal with more than one aspect in the same review sentence. Finally, the proposed model generates a meaningful aspect-based review summary that helps producers and consumers make a decision. The proposed model outperformed the related ones. SALOM is evaluated in terms of Accuracy, Precision, Recall and F-measure respectively.