Chapter 91
Towards Improving the Lexicon-Based Approach for Arabic Sentiment Analysis

Nawaf A. Abdulla
Jordan University of Science and Technology, Jordan

Nizar A. Ahmed
Jordan University of Science and Technology, Jordan

Mohammed A. Shehab
Jordan University of Science and Technology, Jordan

Mahmoud Al-Ayyoub
Jordan University of Science and Technology, Jordan

Mohammed N. Al-Kabi
Zarqa University, Jordan

Saleh Al-rifai
Jordan University of Science and Technology, Jordan

ABSTRACT

The emergence of the Web 2.0 technology generated a massive amount of raw data by enabling Internet users to post their opinions on the web. Processing this raw data to extract useful information can be a very challenging task. An example of important information that can be automatically extracted from the users' posts is their opinions on different issues. This problem of Sentiment Analysis (SA) has been studied well on the English language and two main approaches have been devised: corpus-based and lexicon-based. This work focuses on the later approach due to its various challenges and high potential. The discussions in this paper take the reader through the detailed steps of building the main two components of the lexicon-based SA approach: the lexicon and the SA tool. The experiments show that significant efforts are still needed to reach a satisfactory level of accuracy for the lexicon-based Arabic SA. Nonetheless, they do provide an interesting guide for the researchers in their on-going efforts to improve lexicon-based SA.

DOI: 10.4018/978-1-4666-9840-6.ch091

1. INTRODUCTION

Since the emergence of the Web 2.0 technology, Internet users became capable of sharing their thoughts, views, and comments with the whole world; thus, contributing to the websites contents. Also, rapidly spreading social networks like Twitter, Facebook and Yahoo!-Maktoob encourage such a phenomena. These websites allow Internet users to communicate, debate, and provide their opinions on particular objects. There have been increasing interests over the past years from several parties (including companies and governments) in mining these opinions to obtain useful information about the products or services these parties provide. Subsequently, the field of sentiment analysis has arisen.

Sentiment Analysis (SA) and Opinion Mining (OM) are exchangeable terminologies used for representing the process of automatically extracting the sentiment orientation or polarity of an opinion on a specific object (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). This object can be a person, product, service, event, and so forth. In other words, it determines whether a sentence or a document is positive or negative.[1] These opinions are expressed in various forms such as articles, reviews, forum posts, short comments, tweets, etc.

The benefits of performing SA are countless. SA is essential for companies in our modern life to automatically mine for the perceived advantages/disadvantages of their products/services by the targeted costumers. SA tools can determine the sentiment polarities of thousands of comments on a particular product or service in a very short period of time. By evaluating the sentiments of such comments, the companies can have better plans to improve their products/services; thus, increasing their market share (Pang & Lee, 2008). In addition, SA can also be used by governments to measure the public's opinion on controversial issues as it can serve as a quick and more accurate alternative for public polls. Basically, by analysing what people write on the Internet about a certain issue, the tool can be used to automatically and accurately estimate the public's opinion on this issue.

According to Korayem et al. study (Korayem, Crandall, & Abdul-Mageed, 2012), sentiment analysis studies are classified according to: (I) predicted class (the text is subjective or objective); (II) predicted polarity (be it positive, negative, or neutral); (III) level of classification (SA for a word, phrase, sentence, or a whole document); (IV) the applied approach (supervised or unsupervised). The proposed model in this paper deals with subjective texts. It classifies the whole document (i.e., document-level SA) into one of the three polarity classes (positive, negative or neutral).

SA or OM mainly has two approaches. The first method exploits one or more machine learning classifiers trained on a labelled corpus. After the model construction, it is used to classify the inputted text into one of the predefined classes. This method is called *supervised* or *corpus-based*. On the other hand, the second method depends on a list of words associated with their polarities (+1 or −1), where the model calculates the total polarity of the inputted text from the individual polarities of the words/phrases comprising the inputted text. This method is called *unsupervised* or *lexicon-based*.

Though the first approach proved to produce high accuracy, it has some shortcomings. It requires building a huge corpus (dataset) and labelling it manually by human experts. The process of manual annotation can be very difficult even for native speakers due to sarcasm and cultural references. It can also be expensive and time-consuming (He & Zhou, 2011). Moreover, the model built could be a domain-biased. That is, it could give low accuracy when it is applied on a different domain from which it was learned (Read & Carroll, 2009). On the other hand, in the lexicon-based approach, it requires no enormous, manually annotated corpus. The approach difficulty arises in the lexicon construction and