Chapter 75
Subjective Text Mining for Arabic Social Media

Nourah F. Bin Hathlian
College of Arts and Sciences, Nairiyah University of Hafer Albatin, Alkhbar, Saudi Arabia

Alaaeldin M. Hafez
College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia

ABSTRACT

The need for designing Arabic text mining systems for the use on social media posts is increasingly becoming a significant and attractive research area. It serves and enhances the knowledge needed in various domains. The main focus of this paper is to propose a novel framework combining sentiment analysis with subjective analysis on Arabic social media posts to determine whether people are interested or not interested in a defined subject. For those purposes, text classification methods—including preprocessing and machine learning mechanisms—are applied. Essentially, the performance of the framework is tested using Twitter as a data source, where possible volunteers on a certain subject are identified based on their posted tweets along with their subject-related information. Twitter is considered because of its popularity and its rich content from online microblogging services. The results obtained are very promising with an accuracy of 89%, thereby encouraging further research.

1. INTRODUCTION

Sentiment analysis and classification have become important areas of research related to text mining and natural language processing, particularly in the digital world era. Because of the richness and availability of online sources, several parties, including governmental agencies and private companies, are increasingly relying on those sources to extract information related to their users' opinions and preferences and link it to a broad array of real-world behaviors. The psychological meaning of words: LIWC and computerized text analysis methods, 2010) through the use of sentiment analysis techniques such as natural language processing, computational linguistics, and fundamental text analysis (Haaff, 2010).

DOI: 10.4018/978-1-7998-2460-2.ch075

Various scholars have explored a plethora of sentiment analysis techniques resulting in diverse resources, corpora, and tools available for the implementation of applications like text classification (El-Orfali., 2014) and named entity recognition (Raza, 2009.). However, while these studies are laudable for their insights and contributions, they are limited in their linguistic scope. In other words, most of the studies focused mainly on English texts with few resources available for other languages. Particularly, the Arabic language has received scant attention in sentiment analysis research both at the document and sentence levels (Elhawary & Elfeky, 2010), and there exist very limited annotated resources for sentiment analysis. Consequently, this caused a major bottleneck for applying such techniques to Arabic texts. This is quite surprising, as the Arabic language is one of the ten most used languages on the internet (based on the ranking carried out by the Internet World State in 2010) and is spoken by hundreds of millions of people.

Accordingly, this study attempts to fill this gap by generating a corpus of Arabic text on a particular topic and is therefore classified as sentiment analysis. Specifically, the primary objective is to work on the preprocessing of Arabic tweets to detect the sentiments contained in individuals' opinions and analyze and extract their attitudes. For those purposes, the text will be divided into two classes—interesting and non-interesting—based on a defined subject. The researcher presents the volunteerism concept to classify Arabic text and to predict potential and appropriate volunteers interested in charitable, governmental, or trade organizations through the text mining of their interactions and information on Twitter.

This paper is organized as follows. The second section briefly outlines the existing research in the area of text analysis. The third describes the approach taken in this study and the system's implementation details. The fourth section focuses on the researcher's evaluation of the applied approach. Finally, the last section discusses the conclusion and future research directions.

2. RELATED WORK

Sentiment analysis has evolved as a unique way of text analytics because of the increase of new opinionated data in social media. Sentiment analysis is divided into two main processes: the collection of data sets (corpus) and the categorization of the data depending on their sentiments. In fact, sentiment analysis entails the detection of opinions within the text and the distinction between their polarity classes, whether positive or negative.

Sentiment analysis is performed using a variety of techniques including data mining, natural language processing, and machine learning techniques. Based on previous research, sentiment analysis techniques can generally be classified into two categories. The first category, based on the machine learning approach, represents supervised learning in which a training corpus is initially annotated with its label (either positive or negative), and for each sentence a feature vector is then formed as an input to the machine learning algorithm that creates a classifier model. Consequently, the resulted classifier is automatically built by learning the properties of opinions from a set of training data that is able to predict the classes of new data (Agarwa & Sabharwal, 2012; Abdul-Mageed, 2011). The most common machine learning algorithms are Support Vector Machine (SVM) and Naïve Bayesian (NB). The second category, based on a semantic orientation approach, depends on using a sentiment lexicon of a language. Each sentiment word on the lexicon has polarity weight as a number, which refers to its class (positive, negative, or neutral). The sentence's polarity represents the total of the polarities of its sentiment words determined and extracted from lexicon. Such lexicons are available in English (e.g. SentiWordNet7),