Computers in Human Behavior 54 (2016) 94–100

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Online reputation measurement of companies based on user-generated content in online social networks

Hossein Shad Manaman [a,*], Shahram Jamali [b], Abolfazl AleAhmad [c]

[a] Department of Computer, Khalkhal Branch, Islamic Azad University, Khalkhal, Iran
[b] Computer Engineering Department, University of Mohaghegh Ardabili, Ardabil, Iran
[c] Database Research Group, School of Electrical and Computer Engineering, Campus #2, University of Tehran, North Kargar St., Tehran, Iran

ARTICLE INFO

Article history:
Received 4 February 2015
Received in revised form
17 July 2015
Accepted 30 July 2015
Available online xxx

Keywords:
Twitter
Reputation management
Sentiment analysis
N-Gram classification

ABSTRACT

Social media websites such as Facebook, Twitter, etc. has changed the way peoples communicate and make decision. In this regard, various companies are willing to use these media to raise their reputation. In this paper, a reputation management system is proposed which measures the reputation of a given company by using the social media data, particularly tweets of Twitter. Taking into account the name of the company and its' related tweets, it is determined that a given tweet has either negative or positive impact on the company's reputation or product. The proposed method is based on N-gram learning approach, which consists of two steps: train step and test step. In the training step, we consider four profiles i.e. positive, negative, neutral, and irrelevant profiles for each company. Then 80% of the available tweets are used to build the companies' profiles. Each profile contains the terms that have been appeared in the tweets of each company together with the terms' frequencies. Then in the test step, which is performed on the 20% remaining tweets of the dataset, each tweet is compared with all of the built profiles, based on distance criterion to examine how the given tweet affects a company's reputation. Evaluation of the proposed method indicates that this method has a better efficiency and performance in terms of recall and precision compared to the previous methods such as Neural Network and Bayesian method.

1. Introduction

Information is a valuable component of the world we live in. By the growing size of information available over the Internet, an urgent need is felt towards finding tools which assist in resource search and management. A large share of the generated information content is processed daily over social sites. More and more companies are using social media tools such as Facebook and Twitter to provide various services and interact with customers (Shenghua, Wu, & Ling, 2013). It is well known that each tweet or each post in a social media may not be a representative sample. In other words, it may be a sampling error that can't reflect how the entire population thinks. But, a subset of these statistical samples may accurately reflect the average mentality of the entire population. When, we consider millions of the opinions that are expressed about a certain topic over a social network, the knowledge that is obtained from these social networks, is unlikely to be biased (Mostafa, 2013). In Twitter, people express their comments in the form of Tweets (Internet Messages), which are observable to the public. In this paper, the main focus is on the Twitter system, as the update intervals and the message length are short in Twitter; which consequently demands its specific challenges. Tweets generate a rich source of information on the common public feelings which is valuable for both consumers and organizations.

To make decisions about their purchases, consumers typically rely on the remarks of other people. These decisions are not merely limited to the products; rather, they also cover the services, companies, political nominates, etc. Hence, having a tool which automatically gathers and analysis a large amount of data would be very useful and would have a significant impact on public's life.

Organizations were traditionally doing their validations manually, through the examination and tracing of newspaper articles, which is a grueling and time consuming task. Therefore, there is an urgent need for an automatic tool which accurately detects and classifies the remarks using a huge amount of data. This tool

* Corresponding author.
E-mail addresses: Shadmanaman@qiau.ac.ir (H. Shad Manaman), jamali@iust.ac.ir (S. Jamali), a.aleahmad@ece.ut.ac.ir (A. AleAhmad).

http://dx.doi.org/10.1016/j.chb.2015.07.061
0747-5632/

enables the organizations to recognize the given feedbacks in the early stage of their projects, so that they would have opportunity to correct their mistakes and offer better products and services to consumers. Users spend more time in communities where they have received social-psychological feedback and in communities where they have previously invested more time (Das & Lavoie, 2014). Thus, the remarks and instruments which can gather and give a proper representation of a specific subject or object are useful in both the consumers and commercial aspects. The rest of this paper is organized as follows: Section 2 brings related works and explains previous works in the field. In Section 3, we explain the proposed method. In Section 4, our assessment criteria is described. Finally, in Section 5 we compare and analyze the results of our evaluation.

2. Related works

Sentiment analysis is used to explore subjective information from source materials using natural language processing and text analysis. In other words, sentiment analysis is the process of deriving the attitude of a speaker or a writer with respect to the given topic and detecting the overall contextual polarity of a document. The attitude is based on the evaluation, judgment and emotional state of a speaker or a writer when speaking or writing. Sentiment analysis is implemented to classify the polarity of a given text to determine whether the expressed opinion in the text is positive, negative or neutral. The first research conducted on sentiment analysis was based on manuscripts; in which a sentiment is expressed using a large paragraph of given text used to analyze a particular product (Pang & Vaithyanathan, 2002). Some researchers focus on sentiment analysis using the sentences (Wiebe, Bruce, & O'Hara, 1999), whereas some other concentrate on sentiment analysis based on the comments on a specific feature of a given product or services e.g., battery use of a specific tablet (Hu & Liu, 2004).

Sentiment Analysis is defined as the computational study of opinions, sentiments and emotions expressed in text. Within this broad field, most of the work has been focused on sentiment polarity classification, where a text is classified as having positive or negative sentiment (Ortigosa Hernandez et al., 2012). Sentiment analysis in reviews is the process of exploring product reviews on the internet to determine the overall opinion or feeling about a product (Haddi, Liu, & Shi, 2013). Many concepts used in the mentioned research for sentiment analysis in Twitter are used by other researchers such as Go, Bhayani, & Huange (2009). They considered sentiment analysis as a problem which targets tweets classification on either a negative or positive way. In this regard, standard supervised machine learning algorithms such as Bayesian Classifier and Maximum Entropy were utilized and generated acceptable results; which are similar to the works conducted on movies reviews. However, the Support Vector Machine (SVM) classifier with Unigrams indicate the best performance and accuracy because of its more efficient management of noises and inconsistencies (Pang & Vaithyanathan, 2002).

Some researchers developed this model by using metadata. In addition to the textual content of tweets, information such as user name, hash tags, responses, location, etc. can express the comments of users. Tan, C.,& et al. (2011) have used social connections to create a social graph indicating who is following whom and they show that users who are very 'connected' tend to hold similar opinions, this information along with textual features can better predict the sentiment held by someone. For example, if user 1 and user 2 are two programmers that are strongly connected then if user 1 likes Star Trek, then it is highly probable that user 2 may like Star Trek as well. The users who follow the same people and those who have more communications with each other indicate generally the same opinions. This information coupled with the textural characteristics can predict the sentiments of people. Another researcher applied retweets to understand "user biases" and produce a representing graph about it. They claim that people may change their expressions but their likes are rather unchanged. Therefore, noting this fact can enable the users to have better predictions about the users' sentences (Guerra, Veloso, Meira jr, & Almeida, 2011).

Because of the importance of social communications for understanding the tweets, some researchers used location of people to understanding their sentiments better. These groups of researchers believe that the users of each area apply their own specific expressions and styles during the tweeting, so that incorporation of location as a feature can improve classification accuracy (Davies & Ghahramani, 2011). Along the supervised learning techniques, it seems that SentiWordNet can also be applied as an important resource for sentiment analysis. SentiWordNet lists sentiment score of each word in WordNet, which is a network consisting of a large number of words. Using the sentiment score of each word, researchers analyze online review and blogs and use the results for sentiment recognition. The advantage of this algorithm is that it does not need any training data but it may work in twitter efficiently due to the slang and language style of Twitter users (Khan & Baharudin, 2011).

Amigo, Artiles, & Bing (2010) present a summary of advances and problems of the teams participated in CLEF (2012) in the Online Reputation Management task. Spina, Gonzalo, & Amigo (2011) and Reddy Yerva, Miklos, & Aberer (2012) provide a useful discussion on addressing methods for clarifying the ambiguities in tweets.

3. The proposed method

In our proposed method, first a profile is created for each company to represent its features in different classes. To generate such profile for each class, it is required to apply term selection methods. For this purpose, multivariate chi-square (X^2) method is utilized in this research.

X^2 statistics indicates the participation level of each term in the group. This method is founded on the basis of the assumption that terms with high frequency in a given group are useful for classification purpose. Moreover, to reduce the problem dimensions and avoid the curse of dimensionality, the terms with small X^2 are neglected in this work. In this way, a term matrix which represents the frequency of a term p in a class m is built. Fig. 1 indicates a sample of the matrix.

	e_1		e_k		e_m	
χ^1	N_{11}	...	N_{1k}	...	N_{1m}	$N_{1.}$
…						…
χ^j	N_{j1}	...	N_{jk}	...	N_{jm}	$N_{j.}$
…						…
χ^N	N_{n1}	...	N_{nk}	...	N_{nm}	$N_{n.}$
	$N_{.1}$		$N_{.k}$		$N_{.m}$	$N = N_{..}$

Fig. 1. Matrix of features frequencies in categories (Khan & Baharudin, 2011).