Data Mining in Social Media

Thabit Zatari

Abstract- Social media mining is a process involving the extraction, analysis and representation of useful patterns from data in the social media, deriving from social interactions (Adedoyin-Olowe, Gaber & Stahl, 2013). Social media mining is a young field which has been leading research and development by handling enormous amounts of information. Just like the mining of the minerals, data mining also involve the extraction of useful information from a larger set of data, which is otherwise not evident and is difficult to acquire.

Introduction

An exceptional amount of data is present due to the global use of social media and is of interest to many branches of study such as sociology, business, psychology, entertainment, politics, news and other cultural aspects of societies. We can find compelling aspects of human conduct and human interactions by the application of data mining to social media. we can get a better understanding of the outlook of different people regarding a certain subject, locate groups of people among large communities of people, study changes in group with reference to time, or even suggest a certain product or task to a certain person by using data mining in combination with social media (Barbier & Liu, 2011).

An outstanding example of the use of social media was set during the US president campaign in the elections of 2008. The sites such as twitter and facebook played a very important role in raising finance and delivering the messages of participants to the voters. Thus, the above examples show that the role of data mining in mining the social media data to forecast results at a national level is very significant. We can also fulfill our personal and corporate interests by data mining social media. With a remarkable rise of social media in the recent years, the application of data mining methods to social media data is also getting increasing recognition. The three challenging characteristics of the social media are that it is large, noisy and dynamic. These challenges are overcome by using data mining techniques by the researchers in order to disclose insights into the social media data which is impossible otherwise (Barbier & Liu, 2011).

Various Techniques of data mining and their role in social media

Various data mining techniques have been developed by scientists in order to overcome the problems such as size, noise and dynamic nature of the social media data. Due to the large volume of data in the social media, an automatic data processing is needed in order to analyze it within a given time span. The dynamism in the social media data leads to the rapid evolution of the data sets over time; such dynamic data can be easily handled by various data mining techniques (Adedoyin-Olowe, Gaber, & Stahl, 2013). Different types of data mining techniques are discussed as follows.

1. Unsupervised classification

We can easily evaluate a review as 'thumbs-up' or 'thumbs-down' by using unsupervised learning(75). This type of marking can be done locating the phrases including an adjective or adverb(66). We can imprecise the semantic orientation of every phrase by using PMI-IR (74) followed by the grouping of the review by using the mean semantic orientation of the phrase (Adedoyin-Olowe, Gaber, & Stahl, 2013).

1.1. Sentiment lexicon

Sentiment lexicon can be regarded as a dictionary of the emotional words which are frequently employed by the reviewers in their communication. It comprises a list of ordinary words that helps in the improvement of the data mining techniques when they are used for mining a sentiment in the certain document. Depending upon the diversity in subject matters, various collections of sentiment lexicon can be generated. The sentiment words employed in the sports, for example, are unlike those employed in the politics.We can focus more on topic-specific occurrence by expanding the occurrence of sentiment lexicon combined with the use of high man power (Adedoyin-Olowe, Gaber, & Stahl, 2013).

1.2. Opinion definition and summarization

These are the significant techniques acknowledging opening. Opinion definition can be discovered in a text, sentence or the document's topic, and it can also inhabit the whole document. By analyzing the sentiment polarities, intensity, and affiliated occurrences, opinion summarization aggregates different opinions aired on the particular section of paper.(46). Opinion extraction is crucial for summarization and successive tracking. Using

this technique, the opinionated part is explored in the texts, documents, and topics. It is compulsory to summarize the opinion since all the opinions conveyed in the document are not necessarily of significance regarding the topic under consideration. It plays a very important role in the business organizations and government offices by helping in improving the products and policies respectively (Adedoyin-Olowe, Gaber, & Stahl, 2013).

1.3. Sentiment orientation (SO)

It may be difficult for the potential buyers to make a decision regarding the purchase of a product by tracking usable reviews due to thousands of reviews which are attracted by the widespread products. SO on the other hand is used by the merchants for their ranking standard so that they could avoid the unimportant or confusing reviews presented to the reviewers. The rating is represented in the form of a 5-star scale with 5 denoting the best ranked while one denotes the poor ranking (Adedoyin-Olowe, Gaber, & Stahl, 2013).

1.4. Opinion extraction

This technique is obligatory in order to aim that part of the document comprising authentic opinion. An individual's opinion regarding a specialized subject does not matter unless that specific individual has mastered that specific field. However, the use of both opinion extraction and summarization is mandatory because of the opinion from many people. The larger the number of people giving their opinion regarding a certain subject, the more significant it will be to extract that particular portion (Adedoyin-Olowe, Gaber, & Stahl, 2013).

Other types of unsupervised learning which are being used nowadays are POS (Part of Speech) tagging. Sentiment polarity is the binary classification technique that classifies the opinionated document into predominantly positive or negative opinion (Adedoyin-Olowe, Gaber, & Stahl, 2013).

Bootstrapping is also included in the unsupervised approach which employs available primary classifier to build labeled data upon which a supervised classification can be built upon (Adedoyin-Olowe, Gaber, & Stahl, 2013).

1.5. Basic clustering technique

The small text documentation can be inspected by using k-means and hierarchical agglomerative clustering techniques. The k of the cluster is loaded to the k-means algorithm finding the acknowledged k in the document and in successive documents (Adedoyin-Olowe, Gaber, & Stahl, 2013).

2. Semi-supervised classification

It is a goal-targeted campaign, but it differs from the unsupervised classification in that it can be particularly analyzed. Semi-supervised lexical classification spread the approach to comprise the unlabeled data by integrating lexical knowledge into supervised learning (Adedoyin-Olowe, Gaber, & Stahl, 2013).

Cluster assumption was engaged by making groups of two documents characterized by similar cluster supporting the positive-negative sentiments words as sentiments documents (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3. Supervised classification

A conjunction of multiple bases of facts is used by s supervised learning algorithm in order to mark many adjectives characterized by alike or unlike semantic orientations (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3.1. Support vector machine

It has been regarded as one of extremely tested social media data mining technique. Due to its non-linear nature, this technique easily evaluates the data both theoretically and computationally and has been widely used in sentiment analysis (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3.2. Naïve Bayers

Naïve Bayers counts the occurrences of values and combinations of values in historical data in order to use conditional probabilities. One can also efficiently mine weather forecast by using this technique. It is regarded as one of the three mostly employed supervised learning techniques in the analysis of sentiments (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3.3. Neural Network

Neural networking is a non-linear technique and is regarded as a less widely employed technique of data mining in social media. Neural networking technique is commonly employed for forecasting monetary performance and making decisions regarding business (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3.4. K-nearest Neighbor

K nearest neighbors is an easy algorithm which piles all accessible cases and classifies new cases on the basis of a similarity estimate (e.g., distance functions). KNN is being used in statistical estimation and pattern recognition since the early 1970's as a non-parametric technique (Saedsayad.com, 2015). This technique has not

been very popular regarding the sentiment analysis in the social media (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3.5. Decision tree

This technique, like kNN, has not been regarded as the most used technique in data mining for social media (Adedoyin-Olowe, Gaber, & Stahl, 2013).

3.6. CHAID (Chi-square Automatic Interaction Selection)

One can inspect explicit data by using CHAID classification tool. The dependent variable of two classes is combined by CHAID. The method is used in order to see the respondent's readiness to suggest or not to suggest a certain subject to others on the basis of his own experience regarding the services obtained via patronage (Adedoyin-Olowe, Gaber, & Stahl, 2013)..

3.7. Text mining

The text mining or text data mining is the data mining technique in which one derives high-quality information from texts. Large media companies such as the tribune etc. use the text mining techniques in order to make the information clear and to provide better search experiences to the readers which consequently increases the 'stickiness' of the site and helps the site to generate larger revenue (Karthikeyan&Vyas, 2014).

References

Saedsayad.com,. (2015). Data Mining. Retrieved 1 July 2015, from http://www.saedsayad.com/data_mining.htm

Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617.

Barbier, G., & Liu, H. (2011). Data mining in social media. In Social network data analytics (pp. 327-352). Springer US.

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: an introduction. Cambridge University Press.

Karthikeyan, M., &Vyas, R. (2014).Cloud Computing Infrastructure Development for Chemoinformatics.In Practical Chemoinformatics (pp. 501-528).Springer India.