The Theory and Method of Sentiment Analysis Approaches for Application in the Big Data Frameworks

Jun-jian Tang

ABSTRACT

In some marketing planning or marketing strategy such as consumptive indicator prediction, learning public sentiment in social media have been more and more emphasized in recent years, which can be improved by applications based on Sentiment Analysis in big data. But the theory of Sentiment Analysis in big data has not been well developed. The article studies the problems of the applications of Sentiment Analysis and the suitability of Sentiment Analysis approaches for application in the big data frameworks. Finally the article points out the research difficulties and gives the suggestion about further study and improvement idea.

KEYWORD: Sentiment Analysis, Big Data

INTRODUCTION

The decrease in the cost of both storage and computing power is one of the main factors that led to the booming of big data. Big data enables companies to make targeted, real-time decisions that increase market share. Big data is characterized by the volume, velocity, veracity, variety, value and volatility of data. The distillation and analysis of big data can facilitate a more thorough and insightful understanding of enterprises. SA is typically used to analyse people’s sentiments, opinions, appraisals, attitudes, evaluations and emotions towards such entities as organizations, products, services, individuals, topics, issues, events and their attributes, as presented online via text, video and other means of communication.1

Many techniques for SA can be categorized into the following: Application-oriented, which ranges from stock price predictions to public voice analysis, crowd surveillance and SA-based customer care; fundamental approaches, including word-
level sentiment disambiguation, sentence-level SA, aspect-level SA, concept-level SA, multilingual SA and linguistic features analysis; and social intelligence, which exploits the public’s online content generation to analyses such inputs as pandemic spreading, emotion and responses towards local events.

**SENTIMENT ANALYSIS ISSUES IN BIG DATA**

**Issues in Big Data Analysis**

Big data is associated with the 5 issues, namely volume, velocity, veracity, variety, value and volatility of data. The large amount and high volume of data are the main characteristics of big data. Big data brings not only new data types and storage mechanisms but also new types of analysis. Big data analysis is a continuum and is not an isolated set of activities that involve making “sense” out of large volumes of varied data that, in their raw form, lack a data model to define what each element means in the context of the others.

**Big Data Framework for Sentiment Analysis**

SA mainly focuses on identifying the sentiment of the composer. The approaches to achieve this goal can be divided into two categories, namely content-specific and content-free. Hadoop is useful for pre-processing data to identify macro trends or to find nuggets of information, such as out-of-range values. It enables businesses to unlock potential value from new data using inexpensive commodity servers. MapReduce enables us to take unstructured data, transform (map) such data into something meaningful and then aggregate (reduce) the data for reporting.

Although opinion mining was introduced earlier, SA has gained increasing attention in big data because of the commercial value emphasised by the enterprises (Agnihotri et al., 2015; Harrigan et al., 2014; He et al., 2015). This is because social media is increasingly being relied upon for product reviews. Thus, enterprises have to listen to the voice of the customers online (hence the main advantage that SA offers) and take actions, such as conducting marketing advocacy to promote good feedback about their products, responding to complaints and considering the thoughts of the public in their strategic marketing and product planning. In this aspect, the focus is to understand the sentiment orientation (also known as polarity) of the online message, monitor the sender, as well as understand the topics and themes and the popularity of the message (Batrinca and Treleaven, 2014; Bernabé-Moreno et al., 2015; Malthouse et al., 2013).

**General Approaches of SA**

SA, also known as opinion mining, is the extraction of positive or negative opinions from (unstructured) text (Pang et al., 2002). The idea of mining direction-based text (i.e., text containing opinions, sentiments, affects and biases) was originally proposed by Hearst and Wiebe (Hearst, 1992). In content analysis, traditional forms like topical analysis might not be effective for forums. Therefore, sentiment analysis has recently been used in many forms of web-based discourse (Aggarwal et al., 1997). Sentiment classification has several important characteristics, including various tasks, features and techniques.
Linguistic Rules

Most of the rule-based linguistics approaches are applied to clause-level or concept-level sentiment classification. The algorithm adopts a pure linguistic approach and considers the grammatical dependency structure of the clause by using SA rules. Linguistic rules are useful for dealing with the semantic orientation of context-dependent words (Ding et al., 2007; Sharef and HaghaniKhameneh, 2014) and they are very helpful for extracting implicit features. These features are those that are not clearly mentioned but are rather implied in a sentence.

Sentiment Classification through Machine Learning

The Machine Learning (ML) approach applies the ML algorithm and uses linguistic features with the aim of optimizing the performance of the system using example data. The big data framework such as Mahout and Pentaho contain library and plugins for the ML approach which can be executed to perform the sentiment classification. The text classification methods using the ML approach can be divided into supervised and unsupervised learning methods. The supervised methods use a large number of labeled training documents. The unsupervised methods have high demand because publicly available data are often unlabeled and thus require robust solutions. SVM is a statistical classification method that has been proven to be highly effective method for traditional text categorization compared with other ML techniques (Khairnar and Kinikar, 2013).

Strength/Sentiment Scoring

Sentiment strength is calculated by manipulating the frequency of matched lexicons according to polarity. Such text analysis granularity might still be insufficient, given that a single sentence may contain different opinions about different facets of the same product or service. To this end, concept-level SA (Kontopoulos et al., 2013; Poria et al., 2014a) aims to go beyond a mere word-level analysis of text to provide novel approaches to opinion mining and SA that enable more efficient passage from unstructured textual information to structured machine-processable data in any domain.

APPLICATIONS OF SENTIMENT ANALYSIS

Current states and progress

Recent research indicates that the number of people and companies using social media applications as a customer relationship management tool has dramatically increased (Bagheri et al., 2013; Fuchs et al., 2014; Kaplan and Haenlein, 2010). It is the norm to see a large number of reviews, complaints and compliments posted and shared just seconds after a new product is released. Analysing this information helps companies to accommodate this growing trend in order to achieve some business values like increasing the number of customers; enhancing customer loyalty, customer satisfaction and company reputation; and achieving higher sales and total revenue (Batrinca and Treleaven, 2014; Bravo-Marquez et al., 2014; He et al., 2015).
On the other hand, this information can be used by the customers as testimonials by extracting the strengths and weaknesses of the distinguishable features of each product, as well as finding the satisfaction levels of other users of those products. Applications of SA range from public voice analysis, crowd surveillance, customer care and social intelligence-based SA to exploit the publics’ online content generation for analyzing inputs such as pandemic spreading, emotion and responses towards local events.

**Gaps and Opportunities between Sentiment Analysis Approaches in the Big Data Era**

Although there is increasing awareness and acceptance on utilising big data analytics specifically for SA, as a strategy to improve enterprises’ productivity and profit, it is important to consider whether there is a gap between the big data framework and the SA techniques, so that suitable enhancing studies can be planned. This is mainly because studies in SA have been rooted long before big data frameworks were created and have focused primarily on the content analytics. Existing review-based studies (Medhat et al., 2014; Ravi and Ravi, 2015; Serrano-Guerrero et al., 2015) on SA have focused on the techniques, applications and web services but none of the available studies have focused on the SA approaches’ adaptability for big data. This section intends to discuss whether there are any gaps and suggests future work in this route.

The first point that should be considered is whether the typical approaches in SA are suitable for big data. For this reason, the 5Vs theme in big data is revisited. Several literatures have started to explore the big data issue for SA, such as for the scalability issue (Bing and Chan, 2014; Conejero et al., 2013; Liu et al., 2013), introduction of big data tools for SA (Ding et al., 2013; Mihanović et al., 2014; Prom-on et al., 2014), distributed approach for SA processing (Bravo-Marquez et al., 2014; Fulse et al., 2014; Hossein and Rahnama, 2014) and improved ML models for SA on big data (Bing and Chan, 2014; Ding et al., 2013; Liu et al., 2013; Mukkamala et al., 2014). Undoubtedly, these papers are dated around the year 2014, which marks the booming of the big data era.

In terms of the volume issue, although SA does not specifically concentrate on the amount of data, SA application is expected to work in both small and large scale data. Since SA techniques range from content-specific to content-free approaches, this should not be a problem. On the contrary, the performance of the SA model on a large scale should increase the precision because there are more trainable data; however, the scalability is only studied in depth where the NB classifier is evaluated for scalable SA instead of the standard Mahout library (Liu et al., 2013). However, volume poses a lower influence for SA limitation compared to velocity and variety.

The velocity aspect is closely related in SA because social media is actively used by the users and real-time streaming data is generated. This is the main motivation for the velocity aspect to be studied in several papers (Bravo-Marquez et al., 2014; Kranjc et al., 2014; Xie et al., 2003b; Yu and Wang, 2015b). The velocity issue relates closely with the volume and variety, because the data is generated.
continuously and thus increases the challenge in its analysis. Hence, there is increasing possibility of new linguistic features being created, such as new acronyms, emoticons, idioms and terminologies, which require an update of the SA model. Furthermore, social media messages are by nature shorter and generally not constructed with proper grammatical rules and hence may decrease the text classification accuracy (Bing and Chan, 2014). In this aspect, more advanced SA techniques need to be explored to be able to adapt to the possibility of new linguistic features.

An existing approach based on fuzzy logic has been introduced for opinion mining on large scale twitter data (Bing and Chan, 2014), which was an attempt at mining the meaning of the texts according to the sentiment of the attributes in the text. This method’s performance was also tested in terms of processing time improvement, where the MapReduce framework was used to increase the speed for scanning the texts before the multi-attribute mining. Besides fuzzy logic, a method based on the Hierarchical Dirichlet Process-Latent Dirichlet Allocation (HDP-LDA) was applied for unsupervised aspect identification in the SA. This method also has the ability to automatically determine the number of aspects, distinguish factual words from opinioned words and further effectively extracts the aspect specific sentiment words. The fuzzy logic and LDA approaches have successfully extracted the aspects and meaning, as shown in their experiment results. However, they have been tested on a prepared dataset mainly used for research. In fact, real data generated on social media contains vast amounts of noise. This indicates the need for a capability to sense and identify useful messages from the online media to be used as input for any strategic marketing manoeuvring.

Although SA models are created with an aim to exploit the online social media value, the volatility of the data is going to demand an equal expenditure plan. This is because sometimes the value of the retrieved data is not realised immediately and therefore the issue of how long to store the data requires the attention of both, the data centre officers as well as the strategic planning units. Besides, the pattern of user preferences and behaviour is often described according to temporal features which can be at various intervals according to the customer segmentation profiles. Since generally the data will grow, data management issues such as its storage structure, accessibility control, warehousing and compressing will have to be considered. In this aspect, cloud storage solutions are useful, but only those that feature all these solutions.

CONCLUSIONS

Studies in SA approaches have existed for more than a decade and now are exploited by enterprises as an important tool for strategic marketing planning and manoeuvring. This move is also due to the advancement in data storage, access and analytics enabled through big data frameworks. However, the big data frameworks regard SA as just another possible application that can benefit through its advanced data management. Although several literatures are available that study the challenges of SA in the big data frameworks, such as through the volume, velocity
and variety issue, the value, veracity and volatility have not been explored as much, though in fact taming the data is key for big data analytics. This paper discusses SA approaches and their suitability for the big data framework. The ratio of standard SA approaches to the SA approaches in big data platform is still huge. Implementation and evaluation of the effectiveness of close monitoring of social customer relationship management is also still scarce although big data technologies adoption is healthy. Gaps in the existing approaches and possible future works are suggested according to each of the big data issues. It is predicted that studies and skills development on SA on big data platform for brand monitoring and customer relation management are going to get increasing attention and its growth will be energised by the high demands and a promise of higher revenues for companies. This prediction is supported by analysing the current marketing reports, surveys and summits on SA-based big data analytics for application in customer behaviour understanding and social network comments analysis for consumer sentiments. Furthermore, brand management approaches through SA are expanding and creating a marketing tsunami in many organisations, which has got companies to shift focus towards personalisation and a consumer-centric engagement.

REFERENCES