Journal of Computer Sciences

Literature

Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data

Nurfadhlina Mohd Sharef, Harnani Mat Zin and Samaneh Nadali

Intelligent Computing Research Group, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Article history
Received: 02-09-2015
Revised: 17-02-2016
Accepted: 23-04-2016

Corresponding Author:
Nurfadhlina Mohd Sharef
Intelligent Computing Research Group, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
Email: nurfadhlina@upm.edu.my

Abstract: The ability to exploit public sentiment in social media is increasingly considered as an important tool for market understanding, customer segmentation and stock price prediction for strategic marketing planning and manoeuvring. This evolution of technology adoption is energised by the healthy growth in big data framework, which caused applications based on Sentiment Analysis (SA) in big data to become common for businesses. However, scarce works have studied the gaps of SA application in big data. The contribution of this paper is two-fold: (i) this study reviews the state of the art of SA approaches. including sentiment polarity detection, SA features (explicit and implicit), sentiment classification techniques and applications of SA and (ii) this study reviews the suitability of SA approaches for application in the big data frameworks, as well as highlights the gaps and suggests future works that should be explored. SA studies are predicted to be expanded into approaches that utilise scalability, possess high adaptability for source variation, velocity and veracity to maximise value mining for the benefit of the users.

Keywords: Sentiment Analysis Approaches, Big Data Analytics

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. Prior to this era, companies made decisions based on transactional data stored in relational databases, whereas other potentially important resources in non-traditional and less structured data are ignored. The strategy to leverage big data ranges from evolving current enterprise data architecture to incorporating big data and delivering business value.

Big data enables companies to make targeted, real-time decisions that increase market share. Big data is characterised by the volume, velocity, veracity, variety, value and volatility of data. Nevertheless, the appropriate tools are needed to acquire, organise and derive value from big data to capitalise one hidden relationships and to identify new insights. The distillation and analysis of big data can facilitate a more thorough and insightful understanding of enterprises, which can lead to enhanced productivity, stronger competitive position and greater innovation.

In accordance with the potential that big data offers, an increasing number of studies have focused on techniques for analysing new and diverse digital data streams to reveal new sources of economic value, provide fresh insights into customer behaviour and identify market trends in advance (Bernabé-Moreno *et al*., 2015; Harrigan *et al*., 2014; Malthouse *et al*., 2013). Sentiment Analysis (SA) is one of the main agenda in big data that focuses on various ways to analyse big data to identify patterns and relationships, make informed predictions, deliver actionable intelligence and gain business insight from this steady influx of information.

SA is typically used to analyse people's sentiments, opinions, appraisals, attitudes, evaluations and emotions towards such entities as organisations, products, services, individuals, topics, issues, events and their attributes, as presented online via text, video and other means of communication. These communications can fall into three broad categories, namely positive, neutral and negative. These categories involve many names and slightly different tasks, such as opinion mining, opinion extraction, sentiment mining, subjectivity analysis, customer complaint, affect analysis, emotion analysis, review mining and review analysis.

Many techniques for SA have been introduced. These techniques can be categorised 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 analyse such inputs as pandemic spreading, emotion and responses towards local events. However, no known literature has discussed the issues of SA from the perspective of big data infrastructure, that is, volume, velocity, veracity, variety, value and volatility. This is mainly because in SA, the focus is directed towards content understanding (e.g., polarity, context and content), as opposed to big data infrastructure papers, which highlight the 5 V.

Several papers (Derczynski and Bontcheva, 2014a; Fulse *et al*., 2014; Nirmal and Amalarethinam, 2015a; Xie *et al*., 2003a; Yu and Wang, 2015a) have mentioned that SA on big data is associated with the velocity and volume problem, but a study that reviews the relation between big data issues and SA is unavailable. Existing review-based studies (Medhat *et al*., 2014; Ravi and Ravi, 2015; Serrano-Guerrero *et al*., 2015; Batrinca and Treleaven, 2014) on SA have focused on techniques, applications and web services, but none have focused on the adaptability of SA approaches in big data. This paper addresses this problem and reviews whether the SA techniques, which have been introduced before big data was made popular, are suitable, efficient and effective for big data infrastructure. The main contribution of this paper lies in identifying challenges and making suggestions to solve the gaps.

This paper is organised as follows: The first part briefly introduces SA and its relation to big data. The second part introduces the general issues related to big data. The third part details the approaches of SA, whereas the fourth part describes the future opportunities to solve the issues of SA relation to big data. The conclusion is given in the fifth part.

Sentiment Analysis Issues in Big Data

Although SA is one of the main agenda in big data, no known work has discussed whether SA approaches are suitable for big data infrastructure. This section focuses on this aspect by starting with a discussion of the general scenario and challenges of big data analysis, followed by an exposition about the general SA framework.

Issues in Big Data Analysis

Big data is associated with the 5V 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 and are, in fact, the main reason why the term big data was coined. Having a close relation to volume is the velocity factor, which is related to the process by which real-time streaming data are being generated through sensors and thus need to be analysed. When a huge volume of continuously generated data exists, the veracity issue arises to address the uncertainty, validity, messiness and trustworthiness of the data. The quality and accuracy of the data are also considered, given that these factors are relevant to the variety issue because various formats and styles of data are generated. Next is the issue on the value of the data, which should be exploited promptly. This decision is associated with the volatility or duration in which the data are deemed valid and should thus be stored.

The above facts indicate that 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. Several new issues should be considered when embarking on this new type of analysis; these issues include discovery, iteration, flexible capacity mining and prediction and decision management (Asur and Huberman, 2010; Bravo-Marquez *et al*., 2014; Rao *et al*., 2014).

The discovery issue is attributed to the fact that the value of the data is often hidden deep under the surface of the collected dataset and could only be determined through an exploration process. Furthermore, the actual relationships within the huge amount of data are not always known in advance. Therefore, uncovering insight is often an iterative process until the answers are found. However, the nature of iteration is related to experimentation, such that it sometimes leads down a path that turns out to be a dead end.

An unavoidable issue related to big data is the flexible capacity. Although cloud computing is exploited for big data, the iterative nature of big data analysis requires the utilisation of more time and resources to solve the problems at hand. This challenge is made worse by the fact that big data analysis is not a typical black-and-white decision. Identifying, mining and predicting how the various data elements relate to one another are constant problems. Decision management is also considered in terms of how the implementation of all these actions can be automated and optimised.

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. SA is closely related to opinion mining, which is defined as a quintuple opinion consisting of a target object, feature of the object, a sentiment value of the

opinion, am opinion holder and the time when the opinion is expressed (Sharef and Haghanikhameneh, 2014).

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).

Although studies on SA have progressed over a decade, albeit without emphasis on big data, several platforms provide SA services for big data users owing to its proximity to social media analysis (Batrinca and Treleaven, 2014; Conejero *et al*., 2013; Sharef, 2014) Table 1 shows examples of big data tools. Given the large volume of traffic in social media, the first step in analysing social media is to understand the scope of data that needs to be collected for analysis. Quite often, data can be limited to certain hash tags, accounts and key words.

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. Organisations primarily use Hadoop as a precursor to advanced forms of analytics. Hadoop is a popular choice for filtering, sorting, or pre-processing large amounts of new data in place and distilling such data to generate denser data that theoretically contain more ‘information’. Pre-processing involves filtering new data sources to make them suitable for additional analysis in a data warehouse.

MapReduce enables us to take unstructured data, transform (map) such data into something meaningful and then aggregate (reduce) the data for reporting. All of these steps occur in parallel across all nodes in the Hadoop cluster. A simple example of MapReduce could map social media posts to a list of words and count their occurrences. Such list is then reduced to a count of the number of occurrences of a word per day (Nirmal and Amalarethinam, 2015b).

Once the meaningful data are stored in Hadoop, they can be loaded into an existing enterprise Business Intelligence (BI) platform or analysed directly using powerful self-service tools, such as PowerPivot and PowerView. Customers utilising SQL Server as their enterprise BI platform have a variety of options to access their Hadoop data. These options include Sqoop, SQL Server Integration Services and Polybase.

Oracle has introduced Oracle Advanced Analytics (OAA) to uncover hidden relationships within data by combining in-database algorithms and open-source R algorithms, which are accessible via SQL and R languages. OAA combines high-performance data mining functions with the open-source R language to enable predictive analytics, data mining, text mining, statistical analysis, advanced numerical computations and interactive graphics-all inside the database.

Amazon Web Services (AWS) utilises the AWS Cloud Formation stack, which provides a script for collecting social media messages, such as tweets. The tweets are stored in Amazon S3 and a map per file is customised for use with the Amazon EMR. An Amazon EMR cluster is then created. This cluster uses an SA program within the Python NLTK program, which is implemented with a Hadoop streaming job, to classify the data. The output files are then evaluated to monitor the aggregated sentiment of the tweets.

Big data analytics tools (as shown in Table 2) are mainly characterised by real-time analytics support, which aids users in staying ahead of their competitors. For example, dashboards that draw data from a variety of disjointed systems are developed. These dashboards go beyond a data repository in terms of having many formats (insight) and possessing the ability to construct decisions (actions) based on the tracking of streamed data trends.

The application of SA approaches in analytics tools is mainly driven by companies’ needs for brand management, in which the cycle begins with research on how the company stands in public, followed by an analysis of consumer contents and incorporation of the trends and ingested information into strategic decision-making. Various SA tools can be used to track social marketing. These tools can be classified as either mention analysis or content analysis. The mentioned analysis applications, such as Tweetchup and Sprout Social, do not provide a deep analysis of message contents, but rather report the keyword trend (or hashtags) related to the companies being mentioned in social media. These applications are usually free. Content analysis comes with expensive charges mainly because of its interactive dashboard and Multilanguage ability. These tools include Radian6, Melt water, Simplify 360, Brand watch and Hootsuite, the features of which range from mentions tracking topic analysis and demographics summary. Free applications such as Social Mention also perform content analytics, albeit with low accuracy.

Table 1. Big data tools

Tools	Description
The Hadoop Distributed File System (HDFS)	HDFS divides the data into smaller parts and distributes it across the various servers/nodes. It also enables the underlying storage for the Hadoop cluster.
Server Integration Service (SQL)	These tools allow posts can be downloaded and loaded into Hadoop.
Apache flume	Data can often be gathered for free directly from a social media services public application interfaces, though sometimes there are limitations, or from an aggregation service, such as Data Sift, which pulls many sources together into a standard format.
Map Reduce	Map Reduce is a process that transforms data loaded into Hadoop into a format that can be used for analysis. Map Reduce jobs can be written in a number of programming languages, including. Net, Java, Python, and Ruby, or can be system generated by tools such as Hive (a SQL like language for Hadoop that many data analysts would be immediately comfortable with) provides the interface for the distribution of sub-tasks and the gathering of outputs.
PIG and PIG Latin (Pig and Pig Latin)	Pig programming language is comprised of two key modules: The language itself, called Pig Latin, and the runtime version in which the Pig Latin code is executed. It is configured to assimilate all types of data (structured/unstructured, etc.).
Hive	Hive permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements. It is a runtime Hadoop support rchitecture that leverages Structure Query Language (SQL) with the Hadoop platform.
Jaql	Jaql converts high-level queries into low-level queries and Jaql facilitates parallel processing consisting of Map Reduce tasks. It is a functional, declarative query language designed to process large data sets.
Zookeeper	Zookeeper coordinate parallel processing across big clusters allows a centralized infrastructure with various services, providing synchronization across a cluster of servers.
HBase	HBase is a column-oriented database management system that sits on top of HDFS by using a non-SQL approach.
Cassandra	Cassandra is also a distributed database system. It is designated as a top-level project modeled to handle big data distributed across many utility servers.
Oracle in-database analytics	Include a variety of techniques for finding patterns and relationships in your data. Because these techniques are applied directly within the database, you eliminate data movement to and from other analytical servers, which accelerates information cycle times and reduces total cost of ownership.
Amazon web services	integrates open-source data processing frameworks with the full suite of Amazon Web Services such as Map Reduce, EMR Cluster and NLTK Python

Table 2. Big data analytics tools

Tools	Description
Statistical analysis system	Also known as SAS, it is a software suite developed for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.
Alpine data labs	An advanced analytics interface working with Apache Hadoop and big data with main advantage in terms of collaborative, visual environment to create and deploy analytics workflow and predictive models
Google analytics	Free web analytics service by Google which tracks and reports website traffics
Revolution analytics	Revolution Analytics is the founder of R, an open source and statistical-based software which is useful for statistical computing and graphics. R can be integrated with the Python language which allows efficient programming, a nd MongoDB for scalable data manipulation
Python	A high-level programming language that emphasizes code readability and support multiple programming paradigms.
MongoDB	A storage platform that is a kind of No-SQL database and utilizes JSON-like documents with dynamic formats instead of the traditional table-based relational database
RapidMiner	Open Source environment for machine learning, data mining, text mining, predictive analytics and business analytics.
Mahout	Specifically for machine learning and data mining algorithms using Map Reduce framework, so that the users can reuse them in their data processing without having to rewrite them from the scratch.
Pentaho	Began as a report generating engine but expanded into big data analytics by enabling integration with NoSQL databases such as MongoDB and Cassandra, and Hadoop.
Tableau	A powerful visualization tool that can be integrated with Hadoop Hive to structure the queries and utilizes memory to cache information for interactive data ingestion, manipulation and integration.

Although many these applications have been developed by utilising social media contents, their architecture has not exploited the power of big data ingestion tools. The applications have mainly focused on the crawling and gathering of online messages, classifying the messages for their sentiment categories, extracting subjectivity and customising visualisation. More recent SA applications include Horton Works, which focuses on SA on big data and integrates Flume and Power View to gather and visualise the data. However, this tool has limited SA capabilities because it is only based on the standard sentiment engine in Python NLTK. Only several of the existing SA applications listed above, such as Hootsuite and Radian 6, are based on core SA engines, which include the Alchemy API, Semantria, Lucene and GATE. Applying core SA techniques enables the content analysis to be deeper and more thorough, thus resulting in higher accuracy. These highly specific SA engines are founded by general techniques of SA, which will be discussed in the next section.

General Approaches of SA

Sentiment Polarity Detection

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. In the next sub-sections, we provide a summary of existing methods.

Several tasks are involved in sentiment polarity classification (Banea *et al*., 2014; Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Turney, 2002; Wiebe *et al*., 2005; Wilson *et al*., 2005; Zhuang *et al*., 2006). Three important sentiment polarity tasks are as follows:

- Identifying whether text is objective/subjective or whether subjective text has a positive/negative orientation
- Determining the level of the classification (document/sentence level)
- Identifying the source/target of the sentiment

The two common class problems are concerned with classifying orientation as positive or negative (Pang *et al*., 2002; Turney, 2002). In addition, some researchers worked on classifying messages as opinionated/subjective or factual/objective (Wiebe *et al*., 2004; Wiebe *et al*., 2005). Moreover, some researchers tried to classify emotions, such as happiness, sadness, anger and horror, instead of sentiments (Grefenstette *et al*., 2004; Mishne, 2005; Subasic and Huettner, 2001).

Sentiment polarity classification is classified into document-level, sentence-level and phrase (part of sentence)-level classification. Document-level classification classifies document as positive, negative, or neutral (Mullen and Collier, 2004; Pang *et al*., 2002; Wiebe *et al*., 2005). Sentence-level classification considers and classifies only a sentence (Guo *et al*., 2010; Lee *et al*., 2012), determining whether a sentence is subjective or objective (Riloff *et al*., 2003). To capture multiple sentiments that might exist within a single sentence, phrase-level classification is performed (Wilson *et al*., 2005). Furthermore, to categorise levels and sentiment classes, different assumptions have also been made about sentiment sources and targets (Nasukawa and Yi, 2003). The features and machine learning-based techniques for sentiment polarity classification are detailed in the next section.

SA Features

Explicit Features

In SA studies, four types of explicit features have been used, namely syntactic, semantic, link-based and stylistic features. Syntactic attributes are the most common set of features for SA. Syntactic attributes contain word n-grams (Pang *et al*., 1988; Pang and Lee, 2004), Part-Of-Speech (POS) tags (Gamon, 2004) and punctuation. Moreover, these attributes contain phrase patterns, which make use of POS tag n-gram patterns (Fei *et al*., 2010; Yi *et al*., 2003). They illustrated that phrase patterns like ‘n+aj’ (noun followed by positive adjective) usually denote positive sentiment orientation, whereas ‘n+dj’ (noun followed by negative adjective) often expresses a negative sentiment (Fei *et al*., 2004). In 2004, Wiebe (Bernabé-Moreno et al., 2015) applied collections, where certain parts of fixed n-grams were exchanged with general word tags. Whitelaw *et al*. (2005) applied a set of modifier features (e.g., very, mostly and not). The presence of these features transformed appraisal attributes for lexicon items.

Link/citation analysis is applied in link-based features to detect sentiment from the web and documents. Efron *et al*. (2004) demonstrated that opinion web pages are linked to one another. Link-based features have been used in limited studies. Thus, the effectiveness of such features for SA remains unclear.

Stylistic features contain structural and lexical attributes, which are used in many previous stylometric/authorship works (De Vel *et al*., 2001; Pang *et al*., 1988). Lexical

and structural style markers have been used in limited sentiment analysis studies. Bernabé-Moreno *et al*. (2015) applied hapax legomena (unique/once occurring words) for subjectivity and opinion perception. They found that the presence of unique words in subjective text is higher than in an objective document. Desmet and Hoste (2013) utilised lexical features, such as length of sentence, for the classification of feedback surveys. Lexical style markers (words per message and words per sentence) were used in Cambria *et al*. (2011) to analyse web blogs. Previous studies have shown style markers to be highly common in web discourse (Abbasi, 2005; Zheng *et al*., 2006).

Implicit Features

Studies on implicit features in SA have focused on semantic and linguistic rules to identify the embedded message, which is not typically expressed using predefined keywords. Instead, the meaning is delivered using similar conceptual-based expressions. Semantic features try to identify polarity or provide intensity-related scores to words and phrases. Hatzivassiloglou and McKeown (1997; Bravo-Marquez *et al*., 2014) illustrated a Semantic Orientation (SO) method that was later extended by (Asur and Huberman, 2010). Mutual information was calculated to compute for the SO score of each word/phrase automatically using Turney (Asur and Huberman, 2010).

Moreover, (Rao *et al*., 2014) extended the SO approach using latent semantic analysis. Manual or semi-automatically produced sentiment lexicons (Lee *et al*., 2012; Sharef, 2014; Tong, 2001) commonly use a primary set of automatically generated terms that are manually filtered and coded with polarity and intensity information. User-defined tags are used to indicate whether certain phrases have positive or negative sentiment. Semi-automatic lexicon generation tools were used by (Riloff *et al*., 2003) to construct a set of strong subjectivity, weak subjectivity and objective nouns. They also used other features, such as bag-of-words, to classify English documents as either subjective or objective.

Another method for annotating semantics to words/phrases is Appraisal Group (Zheng *et al*., 2014). Initial term lists are created using WordNet. These lists are then filtered manually to construct the lexicon. Appraisal Theory was developed by (Martin and White, 2005). In this approach, each expression is manually classified into several appraisal classes, such as attitude, polarity of phrases, orientation and graduation. Zheng *et al*. (2014) used Appraisal Group on movie reviews and achieved very good accuracy. Manually generated lexicons have also been used for affect analysis. Subasic and Huettner (2001) applied affect lexicons with fuzzy semantic typing to analyse movie reviews and news articles. Abbasi and Chen, (2007b; 2007a) analysed hate and violence in extremist web forums using manually constructed affect lexicons. Financial index and stock prediction based on SA was explored by (Lee *et al*., 2013; Makrehchi *et al*., 2013; Milea *et al*., 2010; Zhang *et al*., 2011b).

Other semantic attributes contain contextual features that represent the semantic orientation of surrounding text. Semantic attributes have been useful for sentence-level sentiment classification. Subasic and Huettner (2001; Xie *et al*., 2003b) applied semantic features to identify the subjectivity and objectivity of text in a sentence. They also identified the level of subjective and objective clues in a sentence.

WordNet

WordNet was developed in 1986 at Princeton University. It is a large electronic lexical database for English and it continues to be developed and maintained. WordNet consists of synsets from major syntactic categories, such as nouns, verbs, adjectives and adverbs. The current version of WordNet (3.0) contains over 117,000 synsets, comprising over 81,000 noun synsets, 3,600 verb synsets, 19,000 adjective synsets and 3,600 adverb synsets (Poli *et al*., 2010). Most of the current research used WordN*et al*ong with SentiWordNet (Chaumartin *et al*., 2007). WordNet has been used for synonym collection, whereas SentiWordNet has been used to identify the semantic orientation of each sentence or extracted feature.

SentiWordNet

SentiWordNet is a lexical resource for opinion mining. It is a lexicon base that is similar to WordNet, but it is extended with the lexical information about the sentiment of each synset contained in WordNet. Three different polarities, namely positivity, negativity and objectivity, are assigned to each synset in WordNet. The two most common versions of SentiWordNet used in many studies are SentiWordNet 1.0 and SentiWordNet 3.0. Apart from being used in monolingual studies, SentiWordNet can also be used in multilingual SA (Balahur *et al*., 2014; Denecke, 2008; Lim and Kong, 2004; Yong *et al*., 2011).

SenticNet

SenticNet is built by using sentic computing. It is the latest semantic resource specifically developed for concept-level SA. It exploits both Artificial Intelligence (AI) and semantic web technique to recognise, interpret and process natural language opinions better over the web. SenticNet is a knowledge base that can be applied in the development of many fields, such as big social data analysis, human-computer interaction, electronic health and many more (Cambria *et al*., 2011; Poria *et al*., 2014a).

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. All existing works on implicit aspect extraction were based on the use of Implicit Aspect Clue (IAC) and rule-based method to extract implicit aspects. They mapped the implicit aspect to the corresponding explicit aspect (Hai *et al*., 2011; Poria *et al*., 2013; Zeng and Li, 2013).

Sentiment Classification through Machine Learning

The Machine Learning (ML) approach applies the ML algorithm and uses linguistic features with the aim of optimising 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. In the context of big data analysis, a user should determine the type of algorithm that would be applied for the data at hand and such algorithm is executed through big data analytics tools for specific problem-solving purposes, such as predictive analytics.

Typically, two sets of documents are required in an ML-based classification. These documents are the training and testing sets. A training set is used by the classifier to learn the document characteristics, whereas a testing set is used to validate classifier performance.

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 labelled training documents. The unsupervised methods are used when these labelled training documents are difficult to find. The supervised methods achieve reasonable effectiveness but are usually domain specific and language dependent and they require labelled data, which is often labour intensive. Meanwhile, the unsupervised methods have high demand because publicly available data are often unlabelled and thus require robust solutions. Therefore, semi-supervised learning has been introduced and has attracted considerable attention in sentiment classification. In unsupervised learning, it uses a large amount of unlabelled data along with labelled data to build better learning models.

A number of ML techniques have been adopted to perform the classification task in SA (da Silva *et al*., 2014; Go *et al*., 2009; Xia *et al*., 2011). The most popular ML techniques that have achieved great success in text categorisation are Support Vector Machine (SVM), Naive Bayes (NB) and Maximum Entropy (ME). The other well-known ML methods in natural language processing are K-Nearest neighbour, ID3, C5, centroid classifier, winnow classifier and the N-gram model.

Support Vector Machine (SVM)

SVM is a statistical classification method that utilises the structural risk management principle from computational learning theory. SVM has been proven to be highly effective method for traditional text categorisation compared with other ML techniques, such as NB and ME (Khairnar and Kinikar, 2013). SVM also exhibits the best performance for sentiment classification (Prabowo and Thelwall, 2009; Tan and Zhang, 2008; Xia *et al*., 2011; Zhang *et al*., 2011c). When combined with another technique, such as the constrain topic model, SVM is capable of extracting the implicit aspect in reviewed documents (Wang *et al*., 2013a).

Naive Bayes (NB)

NB classifier is a simple probabilistic classifier based on Bayes' theorem. NB is particularly suitable for use when the inputs have high dimensionality. NB is a simple but effective algorithm that has been widely used in document classification works (Ding *et al*., 2007; Melville *et al*., 2009; Tan and Zhang, 2008; Ye *et al*., 2009; Zhang *et al*., 2011a). NB outperforms SVM when the number of features is small (Pang *et al*., 2002). The algorithm also can be improved when combined with other methods, such as senti-lexicon (Kang *et al*., 2012; Sharef and Shafazand, 2014; Zhou *et al*., 2013). A simple NB classifier can be enhanced to enable a better understanding of more complicated models through more appropriate feature selection and unwanted feature (noise) removal (Narayanan *et al*., 2013).

Maximum Entropy (ME)

ME is another ML classifier that has been proven effective in a number of natural language applications. Unlike NB, ME makes no assumptions about the relationship between features, such that it might perform better when conditional independence assumptions are not met. In some cases, such as in the case where words in the lexicon cannot express the sentiment tendency, the ME entropy classification model outperforms lexicon-based methods in terms of identifying sentiment words in a sentence (Fei *et al*., 2010).

Strength/Sentiment Scoring

Sentiment strength is calculated by manipulating the frequency of matched lexicons according to polarity. Extended studies in this challenge include prior polarity (Ghazi *et al*., 2014; Kouloumpis *et al*., 2011; Loia and Senatore, 2013), dependency rules (Poria *et al*., 2014b), negation identification (Wiebe *et al*., 2005) and summarisation (Kontopoulos *et al*., 2013; Zhan *et al*., 2009; Zhuang *et al*., 2006). These approaches, however, are still far from being able to infer the cognitive and affective information associated with natural language, given that they mainly rely on knowledge bases that are still too limited to process text efficiently at the sentence level. Moreover, 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

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. Besides the benefits in entrepreneurship, an analysis of political pages provides information to political parties regarding people's view of their programmes. Social organisations may seek people's opinion on current debates or on matters like the next presidential candidate. This information can be obtained by analysing the sentiment orientation of comments, the number of likes, shares or comments on posted topics.

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 analysing inputs such as pandemic spreading, emotion and responses towards local events. SA that focuses on microblogging is very typical because this is the main source that taps the public's voice. SA on microblogging data is more challenging compared to conventional texts such as documents review, due to the length, repeated use of some unofficial and atypical words and the rapid progress of language variation usage.

For micro blogging SA, especially Twitter, significant work (Cheong and Lee, 2010; Dodds and Harris, 2011; Khan *et al*., 2014; Kontopoulos *et al*., 2013; Sharef and Haghanikhameneh, 2014) has been done through noisy labels, which are also called 'distant supervision'. Twitter is exploited mainly because the nature of the data is textual, compared to the utilisation of Facebook (Eirinaki *et al*., 2012; Ortigosa *et al*., 2014) and YouTube (Cambria *et al*., 2011; Li and Wu, 2010). The social network is also exploited to identify the most influential opinionators (Fukushima *et al*., 2008; Zhao *et al*., 2014) as a communication strategy which is useful during elections and disasters.

Affective computing through SA facilitates answers to questions such as 'What are the important themes that repeatedly feature in user comments?', 'What is the sentiment orientation of a specific gender about a specific post?' and 'What are the trends of happiness and sadness of the user over time?' Emotions in text may be expressed explicitly (for example, emoticons and lexicon) (Fukushima *et al*., 2008; Loia and Senatore, 2013; Ptaszynski *et al*., 2013) as well as implicitly (Balahur *et al*., 2012; Lau *et al*., 2014; Wang *et al*., 2013b). Affective computing enables companies to care more about their customers (Bagheri *et al*., 2013) and is useful for market prediction (Lassen *et al*., 2014; Li and Li, 2013; Milea *et al*., 2012; Nassirtoussi *et al*., 2014; Zhang *et al*., 2009), assists in diagnosing patients' suicidal levels (Desmet and Hoste, 2013; Pestian *et al*., 2010a; 2010b) and allows the related parties to gauge public perception towards events (Loia and Senatore, 2013; Moreo *et al*., 2012). The advancements in affective computing allow applications to sense and deliver services tailored to customer needs, but issues such as privacy need to be observed.

SA has also been tested in multilingual perspectives (Balahur *et al*., 2014; Denecke, 2008; Hogenboom *et al*., 2014; Lim and Kong, 2004; Yong *et al*., 2011) where the focus was to resolve the limitations of language dependent sentiment lexicons. Several approaches exist in this study, such as translating text into a reference language in which a sentiment lexicon is available before subsequently analysing the text and mapping sentiment scores from a semantically enabled reference lexicon to a target lexicon by traversing relations between language-

specific lexicons. These principles have encouraged many languages such as Dutch (Hogenboom *et al*., 2014), Czech (Habernal *et al*., 2014), Malay (Saloot *et al*., 2014) and Arabic (Abdul-Mageed *et al*., 2014) to explore the potential of SA.

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.

Therefore, depending on an ad-hoc or one-off developed model without continuous adaptation and evolving ability might result in limiting the power of the social media analysis. Furthermore, despite the variation of emotion expression and online voice channelling, SA techniques are commonly based on textual sources. In fact, many other multimedia sources should also be processed, some of which are important sources for examples exhibiting expressions of mocking, sabotaging and sarcasm, which are sensitive content for companies' reputations and for competitiveness planning. Therefore, multi-modal SA techniques are probably going to be in high demand in the near future (Fulse *et al*., 2014).

An even more demanding focus is to make sure an SA model stays relevant relating to the veracity issue. This is because besides one work (Derczynski and Bontcheva, 2014b), currently there are very few SA techniques that are able to determine the trustworthiness of the data. Some SA techniques have focused on detecting deceptive reviews and cyber bullying messages (Nadali *et al*., 2013; Shojaee *et al*., 2013) but studies like this are still application-specific. Determining trustworthiness of the data demands more norms and logical reasoning which should be considered using many factors and not limited to only the current message being processed but also other messages being posted by the same message sender, for his profile to be considered. SA techniques should also be updated to be able to reason and determine the levels of uncertainty, validity, messiness and trustworthiness of the data. The quality and accuracy of the developed model must be prioritised. SA algorithms for filtering and pre-processing also have to be updated, to process and consider data which are curated with low control and are possibly meaningless.

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.

Although many analysts and industry experts may suggest that implementers of SA in big data start with small, well-defined projects, learn from each iteration and gradually move on to the next idea or field of inquiry, it is also true that the issues discussed above cannot be subsided to ensure optimum resource utilisation and maximisation of the return on investment.

Conclusion

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.

Acknowledgement

The authors would like to extend their appreciation to Assoc. Prof. Dr. Masrah Azrifah, Dr. Azreen Azman and Assoc. Prof. Dr. Norwati Mustapha for several related research discussions pertaining to the research area.

Funding Information

This research is funded by the Universiti Putra Malaysia, Malaysia.

Author's Contributions

Nurfadhlina Mohd Sharef: Main idea, organisation, preparation of the big data perspectives regarding sentiment analysis.

Harnani Mat Zin and Samaneh Nadali: Sentiment analysis approaches.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abbasi, A., 2005. Applying authorship analysis to extremist-group web forum messages. IEEE Intellig. Syst., 20: 67-75. DOI: 10.1109/MIS.2005.81

Abbasi, A. and H. Chen, 2007a. Affect intensity analysis of dark web forums. Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics, May 23-24, IEEE Xplore Press, New Brunswick, NJ., pp: 282-288. DOI: 10.1109/ISI.2007.379486

Abbasi, A. and H. Chen, 2007b. Analysis of Affect Intensities in Extremist Group Forums. In: Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security, Chen, H., E. Reid, J. Sinai, A. Silke and B. Ganor (Eds.), Springer Science and Business Media, Boston, MA, ISBN-10: 978-0-387-71612-1.

Abdul-Mageed, M., M. Diab and S. Kübler, 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang., 28: 20-37. DOI: 10.1016/j.csl.2013.03.001

Aggarwal, C.C., J.B. Orlin and R.P. Tai, 1997. Optimized crossover for the independent set problem. Operations Res., 45: 226-234. DOI: 10.1287/opre.45.2.226

Agnihotri, R., R. Dingus, M.Y. Hu and M.T. Krush, 2015. Social media: Influencing customer satisfaction in B2B sales. Industrial Market. Manage. DOI: 10.1016/j.indmarman.2015.09.003

Asur, S. and B.A. Huberman, 2010. Predicting the future with social media. Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, Aug. 31 2010-Sept. 3, IEEE Xplore Press, Toronto, ON, pp: 492-499. DOI: 10.1109/WI-IAT.2010.63

Bagheri, A., M. Saraee and F. de Jong, 2013. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Syst., 52: 201-213. DOI: 10.1016/j.knosys.2013.08.011

Balahur, A., J.M. Hermida and A. Montoyo, 2012. Detecting implicit expressions of emotion in text: A comparative analysis. Decision Support Syst., 53: 742-753. DOI: 10.1016/j.dss.2012.05.024

Balahur, A., R. Mihalcea and A. Montoyo, 2014. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. Comput. Speech Lang., 28: 1-6. DOI: 10.1016/j.csl.2013.09.003

Banea, C., R. Mihalcea and J. Wiebe, 2014. Sense-level subjectivity in a multilingual setting. Comput. Speech Lang., 28: 7-19. DOI: 10.1016/j.csl.2013.03.002

Batrinca, B. and P.C. Treleaven, 2014. Social media analytics: A survey of techniques, tools and platforms. Ai Society, 30: 89-116. DOI: 10.1007/s00146-014-0549-4

Bernabé-Moreno, J., A. Tejeda-Lorente, C. Porcel, H. Fujita and E. Herrera-Viedma, 2015. CARESOME: A system to enrich marketing customers acquisition and retention campaigns using social media information. Knowledge-Based Syst., 80: 163-179. DOI: 10.1016/j.knosys.2014.12.033

Bing, L.I. and K.C.C. Chan, 2014. A fuzzy logic approach for opinion mining on large scale twitter data. Proceedings of the ACM 7th International Conference on Utility and Cloud Computing, Dec. 8-11, IEEE Xplore Press, London, pp: 652-657. DOI: 10.1109/UCC.2014.105

Bravo-Marquez, F., M. Mendoza and B. Poblete, 2014. Meta-level sentiment models for big social data analysis. Knowledge-Based Syst., 69: 86-99. DOI: 10.1016/j.knosys.2014.05.016

Cambria, E., M. Grassi, A. Hussain and C. Havasi, 2011. Sentic Computing for social media marketing. Multimedia Tools Applic., 59: 557-577. DOI: 10.1007/s11042-011-0815-0

Chaumartin, F., L. Talana and U. Paris, 2007. UPAR7: A knowledge-based system for headline sentiment tagging. Proceedings of the 4th International Workshop on Semantic Evaluations, (WSE' 07), Stroudsburg, PA, USA, pp: 422-425. DOI: 10.3115/1621474.1621568

Cheong, M. and V.C.S. Lee, 2010. A micro blogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. Inform. Syst. Frontiers, 13: 45-59. DOI: 10.1007/s10796-010-9273-x

Conejero, J., P. Burnap, O. Rana and J. Morgan, 2013. Scaling archived social media data analysis using a hadoop cloud. Proceedings of the IEEE 6th International Conference on Cloud Computing, Jun. 28 Jul. 3, IEEE Xplore Press, Santa Clara, CA, pp: 685-692. DOI: 10.1109/CLOUD.2013.120

Da Silva, N.F.F., E.R. Hruschka and E.R. Hruschka, 2014. Tweet sentiment analysis with classifier ensembles. Decision Support Syst., 66: 170-179. DOI: 10.1016/j.dss.2014.07.003

De Vel, O., A. Anderson, M. Corney and G. Mohay, 2001. Mining e-mail content for author identification forensics. ACM SIGMOD Record, 30: 55-64. DOI: 10.1145/604264.604272

Denecke, K., 2008. Using SentiWordNet for multilingual sentiment analysis. Proceedings of the 24th International Conference on Data Engineering Workshop, Apr. 7-12, IEEE Xplore Press, Cancun, pp: 507-512. DOI: 10.1109/ICDEW.2008.4498370

Derczynski, L. and K. Bontcheva, 2014a. PHEME: Veracity in digital social networks. Proceedings of the Workshop on UMAP Projects Synergy, (UPS' 14), pp: 1-4.

Derczynski, L. and K. Bontcheva, 2014b. PHEME: Veracity in digital social networks. Proceedings of the Workshop on UMAP Projects Synergy, (UPS' 14), pp: 1-4.

Desmet, B. and V. Hoste, 2013. Emotion detection in suicide notes. Expert Syst. Applic., 40: 6351-6358. DOI: 10.1016/j.eswa.2013.05.050

Ding, W., X. Song, L. Guo, Z. Xiong and X. Hu, 2013. A novel hybrid HDP-LDA model for sentiment analysis. Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies, Nov. 17-20, IEEE Xplore Press, Atlanta, GA, pp: 329-336. DOI: 10.1109/WI-IAT.2013.47

Ding, X., B. Liu and S.M. Street, 2007. The utility of linguistic rules in opinion mining. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 23-27, Amsterdam, pp: 811-812. DOI: 10.1145/1277741.1277921

Dodds, P.S. and K.D. Harris, 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PLoS One, 6: e26752-e26752. PMID: 22163266

Efron, M., M. Gary and Z. Julinang, 2004. Cultural orientation: Classifying subjective documents by Cociation analysis. Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music and Design, (AMD' 04), pp: 41-48.

Eirinaki, M., S. Pisal and J. Singh, 2012. Feature-based opinion mining and ranking. J. Comput. Syst. Sci., 78: 1175-1184. DOI: 10.1016/j.jcss.2011.10.007

Fei, X., H. Wang and J. Zhu, 2010. Sentiment word identification using the maximum entropy model. Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering, Aug. 21-23, IEEE Xplore Press, Beijing, pp: 1-4. DOI: 10.1109/NLPKE.2010.5587811

Fei, Z., J. Liu and G. Wu, 2004. Sentiment classification using phrase patterns. Proceedings of the 4th IEEE International Conference on Computer Information Technology, Sept. 14-16, IEEE Xplore Press, pp: 1147-1152. DOI: 10.1109/CIT.2004.1357349

Fuchs, M., W. Höpken and M. Lexhagen, 2014. Big data analytics for knowledge generation in tourism destinations-a case from Sweden. J. Destination Market. Manage., 3: 1-12. DOI: 10.1016/j.jdmm.2014.08.002

Fukushima, Y., F. Masui, M. Ptaszynski and Y. Nakajima, 2008. Macroanalysis of microblogs: An empirical study of communication strategies on twitter during disasters and elections.

Fulse, S., R. Sugandhi and A. Mahajan, 2014. A survey on multimodal sentiment analysis. Int. J. Eng. Res. Technol., 3: 1233-1238.

Gamon, M., 2004. Sentiment classification on customer feedback data: Noisy data, large feature vectors and the role of linguistic analysis. Proceedings of the 20th International Conference on Computational Linguistics, (CCL' 04), Stroudsburg, PA, USA, pp: 1-7. DOI: 10.3115/1220355.1220476

Ghazi, D., D. Inkpen and S. Szpakowicz, 2014. Prior and contextual emotion of words in sentential context. Comput. Speech Lang., 28: 76-92. DOI: 10.1016/j.csl.2013.04.009

Go, A., R. Bhayani and L. Huang, 2009. Twitter sentiment classification using distant supervision.

Grefenstette, G., Y. Qu, J. Shanahan and D. Evans, 2004. Coupling niche browsers and affect analysis for an opinion mining application. Proceedings of the 12th International Conference Recherche d'Information Assistee par Ordinateur, Apr. 26-28, France, pp: 186-194.

Guo, Y., Z. Shao and N. Hua, 2010. Automatic text categorization based on content analysis with cognitive situation models. Inform. Sci., 180: 613-630. DOI: 10.1016/j.ins.2009.11.012

Habernal, I., T. Ptáček and J. Steinberger, 2014. Supervised sentiment analysis in Czech social media. Inform. Process. Manage., 50: 693-707. DOI: 10.1016/j.ipm.2014.05.001

Hai, Z., K. Chang and J. Kim, 2011. Implicit feature identification via co-occurrence association rule mining. Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing, Feb. 20-26, Springer Berlin Heidelberg, pp: 393-404. DOI: 10.1007/978-3-642-19400-9_31

Harrigan, P., G. Soutar, M.M. Choudhury and M. Lowe, 2014. Modelling CRM in a social media age. Aus. Market. J., 23: 27-37. DOI: 10.1016/j.ausmj.2014.11.001

Hatzivassiloglou, V. and K.R. McKeown, 1997. Predicting the semantic orientation of adjectives. Proceedings of the 8th Conference on European chapter of the Association for Computational Linguistics, (ACL' 97), Stroudsburg, PA, USA, pp: 174-181. DOI: 10.3115/976909.979640

He, W., H. Wu, G. Yan, V. Akula and J. Shen, 2015. A novel social media competitive analytics framework with sentiment benchmarks. Inform. Manage., 52: 801-812. DOI: 10.1016/j.im.2015.04.006

Hearst, M., 1992. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th Conference on Computational Linguistics, (CCL' 92), Stroudsburg, PA, USA, pp: 539-545. DOI: 10.3115/992133.992154

Hogenboom, A., B. Heerschop, F. Frasincar, U. Kaymak and F. de Jong, 2014. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Syst., 62: 43-53. DOI: 10.1016/j.dss.2014.03.004

Hossein, A. and A. Rahnama, 2014. Distributed real-time sentiment analysis for big data social streams. Proceedings of the International Conference on Control, Decision and Information Technologies, Nov. 3-5, IEEE Xplore Press, Metz, pp: 789-794. DOI: 10.1109/CoDIT.2014.6996998

Kang, H., S.J. Yoo and D. Han, 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Syst. Applic., 39: 6000-6010. DOI: 10.1016/j.eswa.2011.11.107

Kaplan, A. and M. Haenlein, 2010. Users of the world, unite! The challenges and opportunities of social media. Bus. Horizons, 53: 59-68.

Khairnar, J. and M. Kinikar, 2013. Machine learning algorithms for opinion mining and sentiment classification. Int. J. Sci. Res. Public., 3: 1-6.

Khan, F.H., S. Bashir and U. Qamar, 2014. TOM: Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst., 57: 245-257. DOI: 10.1016/j.dss.2013.09.004

Kontopoulos, E., C. Berberidis, T. Dergiades and N. Bassiliades, 2013. Ontology-based sentiment analysis of twitter posts. Expert Syst. Applic., 40: 4065-4074. DOI: 10.1016/j.eswa.2013.01.001

Kouloumpis, E., T. Wilson and J. Moore, 2011. Twitter sentiment analysis: The Good the Bad and the OMG!. ICWSM.

Kranjc, J., J. Smailović, V. Podpečan, M. Grčar and M. Žnidaršič *et al*., 2014. Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. Inform. Process. Manage., 51: 187-203. DOI: 10.1016/j.ipm.2014.04.001

Lassen, N.B., R. Madsen and R. Vatrapu, 2014. Predicting iPhone sales from iPhone tweets. Proceedings of the 18th International Enterprise Distributed Object Computing Conference, Sept. 1-5, IEEE Xplore Press, Ulm, pp: 81-90. DOI: 10.1109/EDOC.2014.20

Lau, R.Y.K., C. Li and S.S.Y. Liao, 2014. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. Decis. Support Syst., 65: 80-94. DOI: 10.1016/j.dss.2014.05.005

Lee, H., M. Surdeanu, B. Maccartney and D. Jurafsky, 2013. On the importance of text analysis for stock price prediction, AMiner.

Lee, M.C., J.W. Chang, T.C. Hsieh, H.H. Chen and C.H. Chen, 2012. A sentence similarity metric based on semantic patterns. Adv. Inform. Sci. Service Sci., 4: 576-585. DOI: 10.4156/AISS.vol4.issue18.71

Li, N. and D.D. Wu, 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis. Support Syst., 48: 354-368. DOI: 10.1016/j.dss.2009.09.003

Li, Y.M. and T.Y. Li, 2013. Deriving market intelligence from microblogs. Decis. Support Syst., 55: 206-217. DOI: 10.1016/j.dss.2013.01.023

Lim, L. and T.E. Kong, 2004. Building an ontology-based multilingual lexicon for word sense disambiguation in machine translation. Proceedings of the 5th Workshop on Multilingual Lexical Database, (MLD' 04), pp: 1-34.

Liu, B., E. Blasch, Y. Chen, D. Shen and G. Chen, 2013. Scalable sentiment classification for big data analysis using naïve bayes classifier. Proceedings of the International Conference on Big Data, Oct. 6-9, IEEE Xplore Press, Silicon Valley, CA, pp: 99-104. DOI: 10.1109/BigData.2013.6691740

Loia, V. and S. Senatore, 2013. A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content. Knowledge-Based Syst., 58: 75-85. DOI: 10.1016/j.knosys.2013.09.024

Makrehchi, M., S. Shah and W. Liao, 2013. Stock prediction using event-based sentiment analysis. Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies, Nov. 17-20, IEEE Xplore Press, Atlanta, GA, pp: 337-342. DOI: 10.1109/WI-IAT.2013.48

Malthouse, E.C., M. Haenlein, B. Skiera, E. Wege and M. Zhang, 2013. Managing customer relationships in the social media era: Introducing the social CRM house. J. Interactive Market., 27: 270-280. DOI: 10.1016/j.intmar.2013.09.008

Martin, J.R. and P.R.R. White, 2005. The language of evaluation: Appraisal in English. Palgrave MacMillan.

Medhat, W., A. Hassan and H. Korashy, 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J., 5: 1093-1113. DOI: 10.1016/j.asej.2014.04.011

Melville, P., O. Ox and R.D. Lawrence, 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jun. 28-Jul. 01, Paris, France, pp: 1275-1284. DOI: 10.1145/1557019.1557156

Mihanović, A., H. Gabelica and Ž. Krstić, 2014. Big data and sentiment analysis using KNIME: Online reviews Vs. social media. Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, May 26-30, IEEE Xplore Press, Opatija, pp: 1464-1468. DOI: 10.1109/MIPRO.2014.6859797

Milea, V., R. Almeida, N.M. Sharef, U. Kaymak and F. Frasincar, 2012. Computational content analysis of European central bank statements. Int. J. Comput. Inform. Syst. Indust. Manage. Applic., 4: 628-640.

Milea, V., N.M. Sharef, T. Martin, R.J. Almeida and U. Kaymak *et al*., 2010. Prediction of the MSCI euro index based on fuzzy grammar fragments extracted from European central bank statements. Proceedings of the International Conference on Soft Computing and Pattern Recognition, Dec. 7-10, IEEE Xplore Press, Paris, pp: 231-236. DOI: 10.1109/SOCPAR.2010.5686083

Mishne, G., 2005. Experiments with mood classification in blog posts. Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access, (TIA' 05), pp: 1-8.

Moreo, A., M. Romero, J.L. Castro and J.M. Zurita, 2012. Lexicon-based comments-oriented news sentiment analyzer system. Expert Syst. Applic., 39: 9166-9180. DOI: 10.1016/j.eswa.2012.02.057

Mukkamala, R.R., A. Hussain and R. Vatrapu, 2014. Fuzzy-set based sentiment analysis of big social data. Proceedings of the 18th International Enterprise Distributed Object Computing Conference, Sept. 1-5, IEEE Xplore Press, Ulm, pp: 71-80. DOI: 10.1109/EDOC.2014.19

Mullen, T. and N. Collier, 2004. Sentiment analysis using support vector machines with diverse information sources. Proceedings of the EMNLP 9th Conference on Empirical Methods in Natural Language Processing, (NLP' 04), pp: 412-418.

Nadali, S., M.A.A. Murad, N.M. Sharef and A. Mustafa, 2013. A review of cyberbullying detection: An overview. Proceedings of the 13th International Conference on Intelligent Systems Design and Applications, Dec. 8-10, IEEE Xplore Press, Bangi, pp: 326-331. DOI: 10.1109/ISDA.2013.6920758

Narayanan, V., I. Arora and A. Bhatia, 2013. Fast and accurate sentiment classification using an enhanced Naive Bayes model. Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning Oct. 20-23, Hefei, China, pp: 194-201. DOI: 10.1007/978-3-642-41278-3_24

Nassirtoussi, A.K., T.Y. Wah, S.R. Aghabozorgi and D.N.C. Ling, 2014. Text mining for market prediction: A systematic review. Expert Syst. Applic., 41: 7653-7670. DOI: 10.1016/j.eswa.2014.06.009

Nasukawa, T. and J. Yi, 2003. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the International Conference on Knowledge Capture, Oct. 23-25, Sanibel Island, FL, USA, pp: 70-77. DOI: 10.1145/945645.945658

Nirmal, V.J. and D.I.G. Amalarethinam, 2015a. Parallel implementation of big data pre-processing algorithms for sentiment analysis of social networking data. Int. J. Fuzzy Math. Archive, 6: 149-159.

Nirmal, V.J. and D.I.G. Amalarethinam, 2015b. Parallel implementation of big data pre-processing algorithms for sentiment analysis of social networking data. Int. J. Fuzzy Math. Archive, 6: 149-159.

Ortigosa, A., J.M. Martín and R.M. Carro, 2014. Sentiment analysis in Facebook and its application to e-learning. Comput. Hum. Behavior, 31: 527-541. DOI: 10.1016/j.chb.2013.05.024

Pang, B. and L. Lee, 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, (ACL' 04), Stroudsburg, PA, USA, pp: 271-271. DOI: 10.3115/1218955.1218990

Pang, B., L. Lee and S. Vaithyanathan, 1988. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (NLP' 88), Stroudsburg, PA, USA, pp: 79-86. DOI: 10.3115/1118693.1118704

Pang, B., L. Lee, H. Rd and S. Jose, 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (NLP' 02), Stroudsburg, PA, USA, pp: 79-86. DOI: 10.3115/1118693.1118704

Pestian, J., H. Nasrallah, P. Matykiewicz, A. Bennett and A. Leenaars, 2010. Suicide note classification using natural language processing: A content analysis. Biomed. Inform. Insights, 3: 19-28. DOI: 10.4137/BII.S4706

Poli, R., M. Healy and A. Kameas, 2010. Theory and Applications of Ontology: Computer Applications. 1st Edn., Springer Science and Business Media, Springer Verlag, ISBN-10: 9048188474, pp: 576.

Poria, S., E. Cambria, A. Gelbukh and C. Gui, 2014a. A rule-based approach to aspect extraction from product reviews. Proceedings of the 2nd Workshop on Natural Language Processing for Social Media, (PSM' 14), Dublin, Ireland, pp: 38-43.

Poria, S., E. Cambria, G. Winterstein and G.B. Huang, 2014b. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. Knowledge-Based Syst., 69: 45-63. DOI: 10.1016/j.knosys.2014.05.005

Poria, S., A. Gelbukh, A. Hussain, N. Howard and D. Das *et al*., 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intellig. Syst., 28: 31-38.

Prabowo, R. and M. Thelwall, 2009. Sentiment analysis: A combined approach. J. Inform., 3: 143-157. DOI: 10.1016/j.joi.2009.01.003

Prom-on, S., S.N. Ranong and P. Jenviriyakul, 2014. DOM: A big data analytics framework for mining Thai public opinions. Proceedings of the International Conference on Computer, Control, Informatics and Its Applications, Oct. 21-23, IEEE Xplore Press, Bandung, pp: 1-6. DOI: 10.1109/IC3INA.2014.7042591

Ptaszynski, M., R. Rzepka, K. Araki and Y. Momouchi, 2013. Automatically annotating a five-billion-word corpus of Japanese blogs for sentiment and affect analysis. Comput. Speech Lang., 28: 38-55. DOI: 10.1016/j.csl.2013.04.010

Rao, Y., Q. Li, X. Mao and L. Wenyin, 2014. Sentiment topic models for social emotion mining. Inform. Sci., 266: 90-100. DOI: 10.1016/j.ins.2013.12.059

Ravi, K. and V. Ravi, 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Syst., 89: 14-46. DOI: 10.1016/j.knosys.2015.06.015

Riloff, E., J. Wiebe and T. Wilson, 2003. Learning subjective nouns using extraction pattern bootstrapping. Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL, (LLH' 03), Morristown, NJ, USA, pp: 25-32. DOI: 10.3115/1119176.1119180

Saloot, M.A., N. Idris and R. Mahmud, 2014. An architecture for Malay tweet normalization. Inform. Process. Manage., 50: 621-633.

Serrano-Guerrero, J., J.A. Olivas, F.P. Romero and E. Herrera-Viedma, 2015. Sentiment analysis: A review and comparative analysis of web services. Inform. Sci., 311: 18-38. DOI: 10.1016/j.ins.2015.03.040

Sharef, N.M., 2014. Review of sentiment analysis approaches in big data era. Proceedings of the Malaysian National Conference of Databases, (NCD' 14), Serdang, pp: 7-12.

Sharef, N.M. and F. Haghanikhameneh, 2014. Content-Based Analysis Method for Sentiment Scoring in Microblogging Mining. In: New Trends in Software Methodologies, Tools and Techniques, Fujita, H. and S. Corporation, IOS Press, Amsterdam, ISBN-10: 1607506297, pp: 398-414.

Sharef, N.M. and M.Y. Shafazand, 2014. An improved deep learning-based approach for sentiment mining. Proceedings of the 4th World Congress on Information and Communication Technologies, Dec. 8-11, IEEE Xplore Press, Bandar Hilir, pp: 344-348. DOI: 10.1109/WICT.2014.7077291

Shojaee, S., M.A.A. Murad, A. Azman and N.M. Sharef, 2013. Detecting deceptive reviews using lexical and syntactic features. Proceedings of the 13th International Conference on Intelligent Systems Design and Applications, Dec. 8-10, IEEE Xplore Press, Bangi, pp: 53-58. DOI: 10.1109/ISDA.2013.6920707

Subasic, P. and A. Huettner, 2001. Affect analysis of text using fuzzy semantic typing. IEEE Trans. Fuzzy Syst., 9: 483-496. DOI: 10.1109/91.940962

Tan, S. and J. Zhang, 2008. An empirical study of sentiment analysis for Chinese documents. Expert Syst. Applic., 34: 2622-2629. DOI: 10.1016/j.eswa.2007.05.028

Tong, R., 2001. An operational system for detecting and tracking opinions in on-line discussion. Proceedings of the ACM SIGIR Workshop on Operational Text Classification, (OTC' 01), pp: 1-6.

Turney, P.D., 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (ACL' 02), Stroudsburg, PA, USA, pp: 417-424. DOI: 10.3115/1073083.1073153

Turney, P.D. and M.L. Littman, 2003. Measuring praise and criticism. ACM Trans. Inform. Syst., 21: 315-346. DOI: 10.1145/944012.944013

Wang, W., H. Xu and X. Huang, 2013a. Implicit feature detection via a constrained topic model and SVM. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Oct. 18-21, Seattle, Washington, USA, pp: 903-907.

Wang, W., H. Xu and W. Wan, 2013b. Implicit feature identification via hybrid association rule mining. Expert Syst. Applic., 40: 3518-3531. DOI: 10.1016/j.eswa.2012.12.060

Whitelaw, C., N. Garg and S. Argamon, 2005. Using appraisal groups for sentiment analysis. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Oct. 31-Nov. 05, Bremen, Germany, pp: 625-631. DOI: 10.1145/1099554.1099714

Wiebe, J., T. Wilson and C. Cardie, 2005. Annotating expressions of opinions and emotions in language. Lang. Resources Evaluat., 39: 165-210. DOI: 10.1007/s10579-005-7880-9

Wiebe, J., T. Wilson, R. Bruce, M. Bell and M. Martin, 2004. Learning subjective language. Comput. Linguist., 30: 277-308. DOI: 10.1162/0891201041850885

Wilson, T., J. Wiebe and P. Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, (MNL' 05), Stroudsburg, PA, USA, pp: 347-354. DOI: 10.3115/1220575.1220619

Xia, R., C. Zong and S. Li, 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Inform. Sci., 181: 1138-1152. DOI: 10.1016/j.ins.2010.11.023

Xie, Y., Z. Chen, Y. Cheng, K. Zhang and A. Agrawal *et al*., 2003a. Detecting and tracking disease outbreaks by mining social media data. Proceedings of the 23rd International Joint Conference on Artificial Intelligence, (CAI' 3), AAAI Press, pp: 2958-2960.

Xie, Y., Z. Chen, Y. Cheng, K. Zhang and A. Agrawal *et al*., 2003b. Detecting and tracking disease outbreaks by mining social media data. Proceedings of the 23rd International Joint Conference on Artificial Intelligence, (CAI' 3), AAAI Press, pp: 2958-2960.

Ye, Q., Z. Zhang and R. Law, 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Applic., 36: 6527-6535. DOI: 10.1016/j.eswa.2008.07.035

Yi, J., T. Nasukawa, R. Bunescu and W. Niblack, 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. Proceedings of the 3rd IEEE International Conference on Data Mining, Nov. 19-22, IEEE Xplore Press, pp: 427-434. DOI: 10.1109/ICDM.2003.1250949

Yong, K.K., R. Mahmud and C.S. Woo, 2011. Lexical database for multiple languages: Multilingual word semantic network. Eng. Technol., 80: 229-234.

Yu, Y. and X. Wang, 2015a. World cup 2014 in the Twitter world: A big data analysis of sentiments in U.S. sports fans' tweets. Comput. Hum. Behav., 48: 392-400. DOI: 10.1016/j.chb.2015.01.075

Yu, Y. and X. Wang, 2015b. World cup 2014 in the Twitter world: A big data analysis of sentiments in U.S. sports fans' tweets. Comput. Hum. Behav., 48: 392-400. DOI: 10.1016/j.chb.2015.01.075

Zeng, L. and F. Li, 2013. A Classification-Based Approach for Implicit Feature Identification. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, Sun, M., M. Zhang, D. Lin and H. Wang (Eds.), Springer, Berlin, ISBN-10: 3642414915, pp: 190-202.

Zhan, J., H.T. Loh and Y. Liu, 2009. Gather customer concerns from online product reviews-a text summarization approach. Expert Syst. Applic., 36: 2107-2115. DOI: 10.1016/j.eswa.2007.12.039

Zhang, X., H. Fuehres, P.A. Gloor, X. Zhang and P.A.G. Hauke Fuehres, 2009. Predicting stock market indicators through twitter "I hope it is not as bad as I fear. Proc. Soc. Behav. Sci., 26: 55-62. DOI: 10.1016/j.sbspro.2011.10.562

Zhang, W., T. Yoshida and X. Tang, 2011a. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst. Applic., 38: 2758-2765. DOI: 10.1016/j.eswa.2010.08.066

Zhang, X., H. Fuehres and P.A. Gloor, 2011b. Predicting stock market indicators through twitter "I hope it is not as bad as I fear. Proc. Soc. Behav. Sci., 26: 55-62. DOI: 10.1016/j.sbspro.2011.10.562

Zhang, Z., Q. Ye, Z. Zhang and Y. Li, 2011c. Sentiment classification of internet restaurant reviews written in Cantonese. Expert Syst. Applic., 38: 7674-7682. DOI: 10.1016/j.eswa.2010.12.147

Zhao, K., G. Greer, B. Qiu, P. Mitra and K. Portier *et al*., 2014. Finding influential users of online health communities: A new metric based on sentiment influence. J. Am. Med. Inform. Assoc., 21: e212-e218. DOI: 10.1136/amiajnl-2013-002282

Zheng, R., J. Li, H. Chen and Z. Huang, 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. J. Am. Society Inform. Sci. Technol., 57: 378-393. DOI: 10.1002/asi.v57:3

Zheng, X., Z. Lin, X. Wang, K.J. Lin and M. Song, 2014. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowledge-Based Syst., 61: 29-47. DOI: 10.1016/j.knosys.2014.02.003

Zhou, S., Q. Chen and X. Wang, 2013. Active deep learning method for semi-supervised sentiment classification. Neurocomputing, 120: 536-546. DOI: 10.1016/j.neucom.2013.04.017

Zhuang, L., F. Jing and X.Y. Zhu, 2006. Movie review mining and summarization. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Nov. 05-11, Arlington, VA, USA, pp: 43-50. DOI: 10.1145/1183614.1183625