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, an 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 Amalarathinam, 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

<table>
<thead>
<tr>
<th>Tools</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Hadoop Distributed File System (HDFS)</td>
<td>HDFS divides the data into smaller parts and distributes it across the various servers/nodes.</td>
</tr>
<tr>
<td>Apache flume</td>
<td>These tools allow posts can be downloaded and loaded into Hadoop.</td>
</tr>
<tr>
<td>Map Reduce</td>
<td>Map Reduce is a process that transforms data loaded into Hadoop into a format that can be used</td>
</tr>
<tr>
<td>PIG and PIG Latin (Pig and Pig Latin)</td>
<td>Pig programming language is comprised of two key modules: The language itself, called Pig Latin,</td>
</tr>
<tr>
<td>Hive</td>
<td>Hive permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical</td>
</tr>
<tr>
<td>Jaql</td>
<td>Jaql converts high-level queries into low-level queries and Jaql facilitates parallel processing</td>
</tr>
<tr>
<td>Zookeeper</td>
<td>Zookeeper coordinate parallel processing across big clusters allows a centralized infrastructure</td>
</tr>
<tr>
<td>HBase</td>
<td>HBase is a column-oriented database management system that sits on top of HDFS by using a non-</td>
</tr>
<tr>
<td>Cassandra</td>
<td>Cassandra is also a distributed database system. It is designated as a top-level project modeled</td>
</tr>
<tr>
<td>Oracle in-database analytics</td>
<td>Include a variety of techniques for finding patterns and relationships in your data.</td>
</tr>
<tr>
<td>Amazon web services</td>
<td>Integrates open-source data processing frameworks with the full suite of Amazon Web Services</td>
</tr>
</tbody>
</table>

Table 2. Big data analytics tools

<table>
<thead>
<tr>
<th>Tools</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical analysis system</td>
<td>Also known as SAS, it is a software suite developed for advanced analytics, multivariate analyses,</td>
</tr>
<tr>
<td>Alpine data labs</td>
<td>An advanced analytics interface working with Apache Hadoop and big data with main advantage in</td>
</tr>
<tr>
<td>Google analytics</td>
<td>Free web analytics service by Google which tracks and reports website traffic</td>
</tr>
<tr>
<td>Revolution analytics</td>
<td>Revolution Analytics is the founder of R, an open source and statistical-based software which is</td>
</tr>
<tr>
<td>Python</td>
<td>A high-level programming language that emphasizes code readability and support multiple</td>
</tr>
<tr>
<td>MongoDB</td>
<td>A storage platform that is a kind of No-SQL database and utilizes JSON-like documents with</td>
</tr>
<tr>
<td>RapidMiner</td>
<td>Open Source environment for machine learning, data mining, text mining, predictive analytics and</td>
</tr>
<tr>
<td>Mahout</td>
<td>Specifically for machine learning and data mining algorithms using Map Reduce framework, so that</td>
</tr>
<tr>
<td>Pentaho</td>
<td>Began as a report generating engine but expanded into big data analytics by enabling integration</td>
</tr>
<tr>
<td>Tableau</td>
<td>A powerful visualization tool that can be integrated with Hadoop Hive to structure the queries</td>
</tr>
</tbody>
</table>

156
Although many of these approaches have been developed by utilizing 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 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 WordNet along 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 Haghankhameneh, 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; Nassirtoosi 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; Mukamala 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 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 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


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


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


