Received April 24, 2018, accepted June 18, 2018, date of publication June 28, 2018, date of current version July 30, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2851311*

# Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges

**SHAHID SHAYAA[1], NOOR ISMAWATI JAAFAR[2], SHAMSHUL BAHRI[2], AININ SULAIMAN[2], PHOONG SEUK WAI[2], YEONG WAI CHUNG[2], ARSALAN ZAHID PIPRANI[2], AND MOHAMMED ALI AL-GARADI[2]**

[1]Berkshire Media Sdn Bhd, Petaling Jaya 47800, Malaysia
[2]Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya, Kuala Lumpur 50603, Malaysia

Corresponding author: Noor Ismawati Jaafar (isma_jaafar@um.edu.my)

This work was supported by the University of Malaya under Grant PV003-2017.

**ABSTRACT** The development of IoT technologies and the massive admiration and acceptance of social media tools and applications, new doors of opportunity have been opened for using data analytics in gaining meaningful insights from unstructured information. The application of opinion mining and sentiment analysis (OMSA) in the era of big data have been used a useful way in categorizing the opinion into different sentiment and in general evaluating the mood of the public. Moreover, different techniques of OMSA have been developed over the years in different data sets and applied to various experimental settings. In this regard, this paper presents a comprehensive systematic literature review, aims to discuss both technical aspect of OMSA (techniques and types) and non-technical aspect in the form of application areas are discussed. Furthermore, this paper also highlighted both technical aspects of OMSA in the form of challenges in the development of its technique and non-technical challenges mainly based on its application. These challenges are presented as a future direction for research.

**INDEX TERMS** Opinion mining, sentiment analysis, big data, applications, opinionated data, social media, online social network.

## I. INTRODUCTION

The 21$^{st}$ century has witnessed a torrential flow of data. The data has sprung massively in various fields over the last two decades, which has led to the birth of big data [1]. Moreover, the influx of technology in the digital world has opened the doors for the development of big data. Citizens of the world are now becoming technology savvy with devices ranges from digital sensors, communication tools including social media applications, and actuators and data processors [2]. For instance, organizations capture the mushrooming volume of transactional data, through which trillions of bytes of information are generated regarding aspects from suppliers to customers. The physical world has millions of network sensors embedded in devices like smart phones, smart energy meters, automobiles, and industrial machines [3]. Such advances in digital sensors and communication technologies have led to the development of the Internet of Things (IoT) [4]. With such a development, social networking sites and communication devices like smart phones, laptops, and PCs allow individuals to interact with one another to create massive amounts of big data [3]. For instance, Twitter's wide network of 467 million users generates 175 million tweets on a daily basis [5]. Similarly, the amount of space needed to store one second of a high definition video is 2000 times more than the space needed to store a page of plain text [3]. Furthermore, according to the International Data Corporation report in 2011, the world is already generated about 1 zettabyte (ZB) of data, and the rate at which this amount is growing has been exploding; the amount of data grew to 7ZB by the end of 2014 [6]. Moreover, by 2020, the amount of data generated is expected to reach 44ZB, with at least half of them being textual data [7] that is generated through social media technologies like Facebook, Twitter, and mobile instant messaging apps such as WhatsApp and Telegram. It has been determined that 500 million tweets are sent each day, while 40 million of those are shared daily. Meanwhile, it is estimated that 4.3 billion messages on Facebook are posted with 5.75 billion likes on a daily basis. Moreover, it is expected that the amount of data will continuously grow because of the influx of digital technologies that have already sprung up in the digital era [1].

The extensive use of technologies and astounding flow of data over the years has also aided in the escalation of big data business analytics. As the term suggests, it refers to two components: big data and business analytics. According to the McKinsey Global Institute, "Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" [3]. However, big data is generally defined through the key characteristics of volume, variety, and velocity [8].

1. Volume represents the quantity of data that utilizes massive storage space or entails sizeable and considerable number of records [9]. For instance, Wal-Mart manages to store 2.5 petabytes of information, Tesco generated 1.5 billion new items of data on a monthly basis, and Dell developed a database which can handle 1.5 million sales and advertisement records [3], [10].
2. Variety refers to the data generated from a range of sources and in varying formats [9]. The sources can be in the form of sensors, social media sites, web technologies, mobile phones, etc. The data format can include web logs, unstructured data like audio, videos, images and sensory data from RFID devices, or other smart sensors.
3. Velocity indicates the frequency with which data is produced from different sources [9]. The data can be generated occasionally, frequently, and/or on a real-time basis.

On the other hand, analytics refers to the capability of the firm in applying statistics, econometrics, mathematical, simulation, and optimization tools [11] to obtain insights from data and employ data-driven decisions in the organizations [12]. Hence, companies in the majority of sectors are now focusing on acquiring profound information from the massive quantity of already-available data to gain competitive advantage [13].

Big data offers great challenges to organizations because of the nature of its complexity. At a fundamental level, organizations face the challenge of handling and storing a gigantic amount of data [14]. Moreover, there is a need to improve method for dealing confounding amount of raw data in a variety of forms [15]. Additionally, it is essential to design scalable data storage in order to acquire and retrieve meaningful information efficiently. High speed networking infrastructure can also consume less power when processing the data [2]. Furthermore, distributing information across many systems is another challenge which is critical in processing huge amounts of data from different datasets in a reasonable period of time. Cloud computing programming is one way to overcome this issue [16]. Apart from all these issues, privacy, security, accessibility, and deployment are additional issues that decision makers must consider before exploiting the vast benefits of big data to gain competitive advantage [14].

We commonly think about what other people think in our decision-making process [1]. Prior to the advent of the Internet, many of us relied on friends and families for product or service recommendations, voting views during local elections, or information when buying a product.

The Internet eases our efforts to get opinions and experiences of those that are neither on our personal contact list nor in our professional networks. The amount of opinions and comments on the Internet has not only risen during the last decade or so, but they are also readily available to the strangers. However, opinion mining and sentiment analysis face challenges such as making query classification, classifying documents that contain reviews or explicitly opinionated material, determining sentiments based on the content, and finally presenting the sentiment information in a reasonable summary [1].

The main aims of this paper are as follows:

- ➢ To provide a comprehensive systematic literature review and discuss both the technical aspects of OMSA (techniques, types) and non-technical aspects in the form of application areas.
- ➢ To highlight both technical (related to the development of sentiment technique) and non- technical challenges (related to the application of Sentiment analysis). These challenges are presented as future research direction.

### A. AN OVERVIEW OF TEXT MINING AND ANALYTICS

Since an enormous amount of data has emerged over the years at a staggering rate, there is a need to incorporate some sort of analytics to gain meaningful insights from the raw and unstructured data in the form of text, images, and videos. Text mining is one of the approaches that are the predecessors to text analytics. Text mining uses natural language processing, knowledge management, data mining, and machine learning techniques to process text documents [17]. Text analytics, while similar to text mining in terms of method, usually deal with a bigger amount of data to extract and generate useful non-trivial information and knowledge [18].

Text mining/analytics are originally conducted for two purposes. The first purpose is to analyze people's sentiment on an issue or phenomenon. Hence, sentiment analysis goes through a huge amount of textual data to identify people's attitudes, thoughts, judgments, and emotions on an issue [19], [20]. The second purpose is to assess people's opinion on a product, person, event, organization, or topic from a user or group of user perspectives. Similar to sentiment analysis, opinion mining is a natural language processing task that employs an algorithmic technique to recognize opinionated content and categorize it into positive, negative, or neutral polarity [21]. Nonetheless, the application of text mining/analytics has been extended to other areas of human computer applications, and the applications are growing with the growth in big data analytics.

### B. AN OVERVIEW OF OPINION MINING AND SENTIMENT ANALYSIS

An opinion refers to a person's or group's sentiment or views, emotions, and attitudes about a product, service, occasion, or other topic present in the environment. Like sentiment analysis, opinion mining is also grounded on the algorithmic technique [21]. Covering a huge variety of pubic opinions, [22] have argued that opinion can be classified into three main types: regular opinions, which refer to a single entity only; comparative opinions, which compare or contrast more than one entity; and suggestive opinions, which suggest a single or multiple entities. The regular opinion is mainly used to identify a positive or negative outlook of a particular product [23]. On the other hand, comparative opinions help in elucidating the association among multiple entities and are mainly used for competitive intelligence [23]. However, there is a dearth of literature concerning the identification of comparative sentences that is being used for the comparison of multiple entities. Recently, suggestive review has been introduced in the field of opinion mining [24]. The extraction of these types of opinions from text can be utilized for various application areas in the field of business, engineering, medical science, and e-learning. It can be offered to various online communities for their assistance as well [22].

Similarly, private statements of individuals are called sentiments, which comprise thoughts, opinions, attitudes, views, judgments, and feelings. These are commonly gathered by conventional scientific methods [19], [20]. [25] pronounced the feelings that are expressed in language by using subjective expression. The sentiments can be analyzed through the machine learning technique, which can be further classified into supervised and unsupervised, using a lexicon-based approach, using a keyword, and using a concept-based technique [26]. Recently, research on sentiment analysis has focused on multiple modalities such as in speech and video as opposed to earlier work that focused on unimodality related to text [27], [28]. Sentiment analysis tackles many NLP subtasks, including aspect extraction [29], subjectivity detection [30], named entity recognition, and sarcasm detection [31].

In most cases, the main objective of sentiment analysis is to unearth people's opinions to gain meaningful insight about products or services. Its aim is to exhibit useful information to both customers and manufacturers. It is established that both manufacturer and customers look upon summarized opinions instead of detailed reviews. Hence the opinions that are categorized on positive, negative, or neutral sentiments are useful for both parties in making the right call [32]. Despite the large number of studies on opinion mining and sentiment analysis techniques, the impact they have on people has been less explored. There has been great emphasis on the techniques used and less on how people can benefit from the findings. Hence, this study aims to investigate the human element in opinion mining and sentiment analysis research. To achieve this aim, we will systematically review the relevant literatures that have employed both approaches.

The study offers several contributions. The first and foremost significance of this study is to refocus the study of opinion mining and sentiment analysis to both technical and non-technical challenges. Secondly, it places emphasis on the areas of opportunity by looking at the trends of application coverage that would offer some potential areas for research. Thirdly, the paper presents information on different datasets that were used in opinion mining and sentiment analysis studies that future researchers could use in their research.

The remaining section of the paper is organized as follows. First, the study talks about the method employed to achieve the research objectives. Then we present the findings of the study. Next, the paper highlights the commonly used dataset in literature. Later sections discuss different review methods for sentiment analysis and highlight the application area for sentiment analysis. Finally, the study emphasizes the current open challenges of big data sentiment analysis.

## II. METHODOLOGY

The main goal of this paper is to develop a deep understanding of the various opinion mining and sentiment analysis approaches performed on human applications of text analytics. This study advocates the ways applications are present and utilized in many areas in the society. The technique used in the study is the systematic literature review. A systematic review is completely based on an evidently framed question, presents relevant studies, evaluates their findings, and summarizes the evidence by means of clear methodology. This unambiguous and methodical approach makes systematic reviews different from the traditional reviews.

## III. REVIEW OF OPINION MINING AND SENTIMENT ANALYSIS

Literature reviews are rooted in medical science, which has been categorized as a critical approach mainly applied where evidence is important [2]. They involve a stringent process of finding, selecting, and examining secondary data [33], [34]. The synthesis of evidence from the current literature can create new knowledge in the current studies, which is as significant as conducting new research [35].

Rousseau *et al.* [36] maintained an argument that systematic reviews are different from tradition reviews in that systematic literature reviews are comprehensive in nature, use transparent and unbiased analysis, and apply certain criteria for interpretation of the findings that provided in the previous literature. In addition, systematic literature reviews mainly focus on objectivity and reproducibility of results [34]. The process of review starts with framing the questions and conducting a systematic and step-by-step process and applying a replicable method to answer these questions [34]. Thus, the evidence generated from the rigorous approach of identifying, selecting, and analyzing the data can have a significant impact on the body of knowledge collected, but the supreme concern of this practice is synthesizing the results produced through this systematic process [33], [34].

The methodology used in the study is a five-step process shown in Figure 1, as proposed by [33]. It is systematic in nature, clear and reproducible, and involves identifying, examining, synthesizing, deducing, and reporting the evidence from the existing sentiment analysis and opinion mining literature.

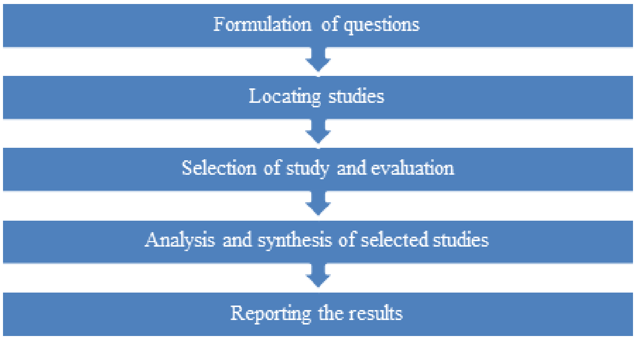


**FIGURE 1.** Research methodology of systematic literature review.

### A. QUESTION FORMULATION

A deep and insightful literature review should start with the development of a clear understanding about your objectives [35]. Therefore, to ascertain this, we clearly formulated and considered research questions to evade doubts in our study [36]. The purpose of the paper is to discuss the methodological and application side of opinion mining and sentiment analysis and explore whether the intervention of opinion mining and sentiment analysis would be applicable to humans or in an organization as a whole. Hence the purpose of our systematic review is to answer two research questions:

*Q1:* What are the trends of opinion mining and sentiment analysis publication from 2000-2016?

*Q2:* Which are the areas in which opinion mining and sentiment analysis have been applied?

*Q3:* What are the sources of data in the areas in which opinion mining and sentiment analysis have been applied?

### B. LOCATING STUDIES

The objective of identifying an appropriate journal articles is to develop a list of all related articles to our research questions. We have selected Web of Science as a core database. The study has only focused on peer-reviewed articles that were written in English and published in the Web of Science Journal category. The study did not consider articles in other categories like conference proceeding papers, book chapters, review papers, and theses.

Since the study is based on opinion mining and sentiment analysis, we used different strings to identify relevant papers. The researcher employed different techniques for searching, including separate keywords for "sentiment analysis" and "opinion mining", combining two keywords at the same time through simple operators and Boolean logic. For example, we used the string "opinion mining" and "sentiment analysis" for exploring the papers containing the exact phrases "opinion mining" and "sentiment analysis."

### C. STUDY SELECTION AND EVALUATION

In order to ensure and maintain the quality of the paper, we have constrained our selection of articles to only peer-reviewed journals. Peer-reviewed journals have strict quality control and have gone through systematic, rigorous processes and have stringent requirements for publication, which leads to better research output [35]. The process began with scanning of selected articles from the Web of Science database. The timeline we set was from the years 2000 to 2016. The initial criterion of selection was based on choosing the keywords "opinion mining" and "sentiment analysis."

We have further followed the criteria suggested by [37] for the inclusion and exclusion of the articles from the selected list. Basically, we have developed the inclusion criterion as follows:

- Published in peer-reviewed journals
- Within the database of the last 15 years (2001 to 2016).
- In the English language
- Containing at least one keyword

The exclusion criteria are as follows:

- Have very narrow horizon or context
- Do not explicitly focus on application side (human or organizational level) of opinion mining and sentiment analysis

Our search resulted in an initial list of 274 articles for sentiment analysis and 91 articles for opinion mining. Thus, the preliminary results provided a total of 365 articles that were written on opinion mining and sentiment analysis.

The next step involved reading the abstracts to evaluate whether it was relevant to our research topic. Initially a single person read it, but to warrant its rigor, an independent person to improve its objectivity and validity read the same number of articles. Scholarly outputs that did not align with our developed research questions or that seemed irrelevant and non-substantive were excluded. The articles that were included exhibited good fit with the objective of the study. Thus, a total of 99 articles were shortlisted based on the initial evaluation.

In the next step, two authors reviewed the pre-selected articles separately. The total number of peer-reviewed journal papers selected for critical reviews after rigorous assessment were 58, published over a period of 16 years. The selected papers were then examined in detail and synthesized to answer the research questions.

For the selected papers, we created the taxonomy which is represented in following sections (dataset, methods, application, and major challenges).

## IV. DATASETS

A more in-depth analysis was done regarding the sources of datasets which are shown in Figure 2. The main source of data is Twitter with seventeen (17) articles, followed by movie reviews with eight (8) articles, Amazon with six (6) articles, blogs with five (5) articles, media with four (4), YouTube with three (3), and Tripadvisor with three (3) articles. There are also nineteen (19) articles that can be categorized as others with one (1) article from each.

## V. METHODS

In this section we discuss the general process for sentiment analysis. A common process of sentiment analysis loop starts

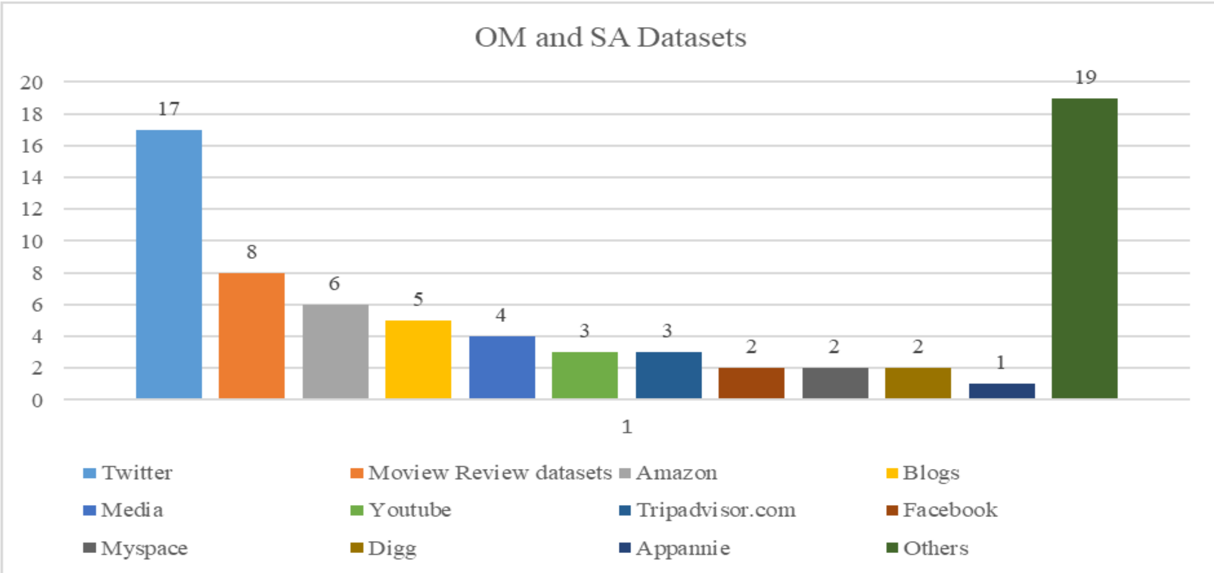


**FIGURE 2.** Opinion mining and sentiment analysis data sets.

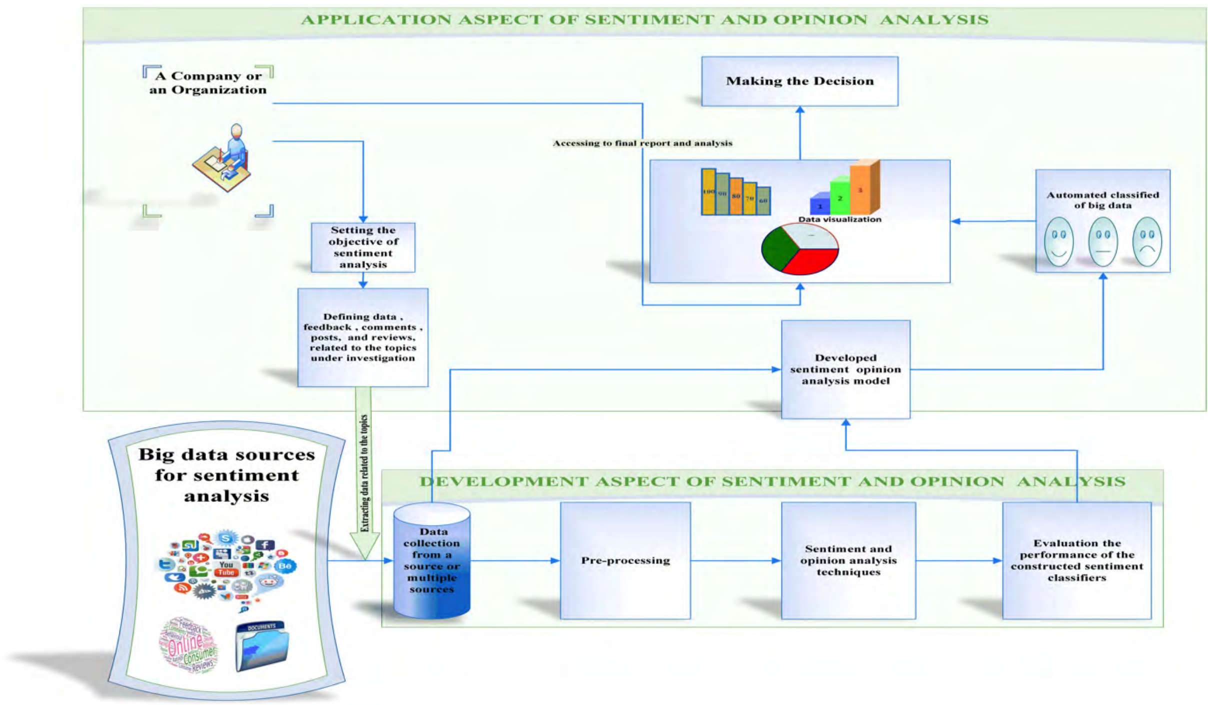


**FIGURE 3.** General process for sentiment analysis.

with goal setting for employing the sentiment analysis. This depends on the application of sentiment analysis (see application section), and then big data is extracted either from a single or multiple sources. The next step is the application of specific sentiment analysis methods in order to mine this huge data and get insight into the company in making final decisions regarding their products. The general sentiment analysis process is shown in Figure 3. In the following subsection, we discuss the most commonly used method to analyze the extracted big data for sentiment analysis. First of all, the main process originates from the collection of the data from data sources such as social media. It should be kept in mind that the collected data should be relevant to the objective of the sentiment system. For example, if the company wants to collect their products, then the data should be extracted with keywords that represent the products. This also involves setting the objective of the sentiment analysis to explain the purpose of sentiment analysis in order to determine the related keywords. The next stage is to pre-process the data in order to remove noise and unrelated content. It should be followed by the construction and evaluation of the sentiment analysis model (based on keywords, lexicon, or machine learning methods). As shown in Figure 4, machine learning algorithms methods commonly involve supervised machine learning methods which require training data to train the algorithms. After building the sentiments and evaluating them in real data sets (test data), the constructed model is applied to unknown big data to automatically classify the methods. Sentiment analysis models (based on keywords method, lexicon or based on machine learning method) are explained in the following subsections.

### A. KEYWORD-BASED CLASSIFICATION

This method classifies text based on the presence of positive or negative polarity words such as happy, joyful, delighted, miserable, sad, terrified, and uninterested [38]. The main drawback of keyword-based classification is the inability to steadfastly classify the negated words and polarity, as this approach depends on surface features [38]. Another drawback is that this approach is based on the obvious presence of positive or negative polarity. However, occasionally, a post may covey sentiment or opinion through underlying meaning rather that obvious polarity words [38].

### B. LEXICON-BASED CLASSIFICATION

Lexicon-based approaches construct lists of words manually labeled as having positive and negative polarity, and a polarity score for each word is created. This constructed lexicon is used to calculate the overall sentiment score of a given post or text. The notable advantage of the lexicon-based method is that these methods do not need training data (as the supervised machine-learning method does). The lexicon-based method is widely used in conventional text like reviews, forums, and blogs [39], [40]. However, they are less likely to be used for big data extracted from social media websites [40]. The key reason is the unstructured format and nature of social media websites (the data contains textual peculiarities, informal and dynamic nature of language, new slang, abbreviations, and new expressions) [40]. Even though this approach outperforms the keyword-based classification, it still has drawbacks. Since it works at the word level, negated posts and posts with other meanings trick the lexicon polarity score measurement [38]. Second, lexical dictionary and polarity scores are usually biased toward the text of a specific type, dictated by the linguistic corpora source [38]. Therefore, it is challenging to construct a more generalized model regardless of the application domain.

### C. MACHINE LEARNING-BASED APPROACH

Machine learning research has become a significant task in numerous application areas. Machine learning reaches throughout recent years have magnificently created algorithms for handling volumes of data to unravel real-world issues. Machine learning algorithms are grouped into supervised learning and unsupervised learning algorithms. Supervised learning algorithms will help users train and learn from the training example which is then tested and evaluated using the test data. The main drawback of supervised machine learning algorithms is the obligation to create a training example. The training example must be comprehensive enough to make the algorithm effective and reliable enough to classify the instance in test data. Another type of machine learning is unsupervised learning algorithms. The working principle of this algorithm is to identify the hidden associations in unlabeled data. The unsupervised learning methods are based on calculating similarity differences between data. For example, it calculates k-means in which similarity between data is computed based on proximity measures, such as Euclidean distance.

The constructing machine learning-based method involves the following important steps.

#### 1) FEATURES EXTRACTION

As shown in Figure 4, feature extraction is an important part of building an effective machine learning method in which the textual posts ($P_1, P_2, P_3, \ldots, P_n$) are transformed into valuable word features ($w_{f1}, w_{f2}, w_{f3}, \ldots w_{fn}$) by using various feature engineering approaches. Feature extracting is one the most important steps of constructing effective classifiers [41]–[44]. The accomplishment or failure of the sentiment classification model is intensely dependent on the features quality. If the extracted features relate well with the sentiment polarity and can provide discrimination power between positive and negative, then classification will be more precise. In contrast, if the extracted features do not relate well with the sentiment polarity and the similar features exist in both positive and negative posts, then the classification task will be more challenging and less precise. The most commonly used features are the first automatic generated features of Bag of Words (BoW), Bag of Phrases (BoP), *n*-gram, and Bag of Concepts (BoC). The second group of features is based on lexical features such as opinion words, sentiment words, and negation words. The third group of features is varied based on the data source; for example, the data from social media normally adds features such as the number of hashtags and social media-related features such as abbreviations and emojis.

#### 2) MACHINE LEARNING ALGORITHMS

This subsection briefly describes the most commonly used machine learning algorithms in literature for sentiment analysis.

##### *a: ARTIFICIAL NEURAL NETWORK (ANN)*

The Artificial Neural Network (ANN) is a mathematical modeling approach that is stimulated by the operative processes of the human mind [45]. It is based on the artificial adaptive system which has some form of distributive architecture [46].

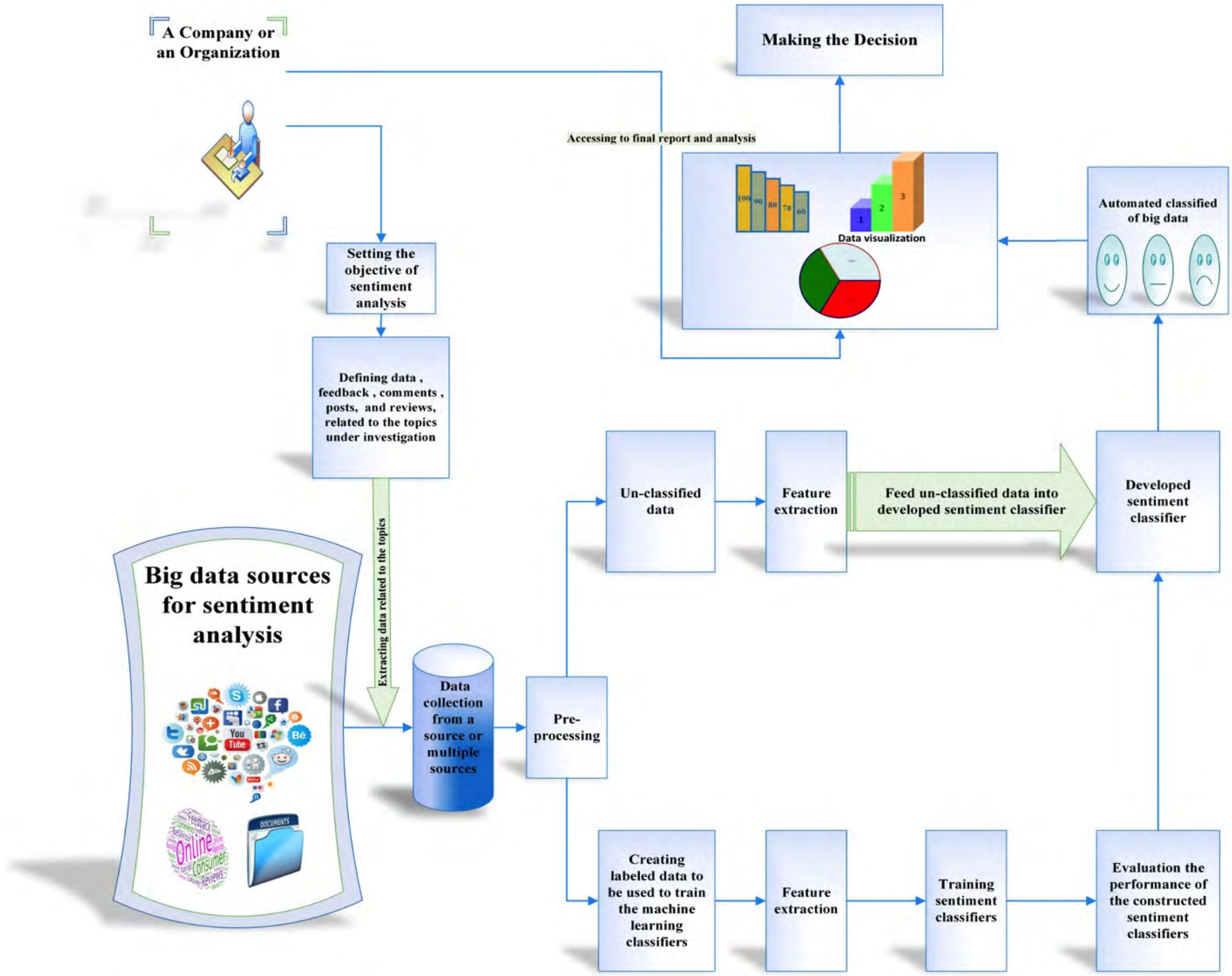


**FIGURE 4.** Steps for constructing machine learning based method.

The system in ANN encompasses closely knit adaptive processing elements called artificial neurons or nodes which are proficient in carrying out enormous analogous and parallel computations for the purpose of information processing and knowledge representation [47].

These systems can also adapt to their internal structure in relation to a functional purpose. There are several types of artificial neural networks which have been used for different problems. Generally, their application is appropriate for a problem which is nonlinear in nature [47]. The nature of the problem can be carrying out pattern recognition, modeling memory, and envisaging the development of a dynamic system [46]. In most cases, these network types perform data modeling which is driven in a supervised or unsupervised fashion. The supervised learning technique is provided with both input and output. Then the network uses this input to generate output, and hence it is compared to desired outputs. On the other hand, in unsupervised training, only input is being provided and the network is used to find natural grouping with a dataset independent of external constraints. That is, the system itself must have the capability to decide which features to incorporate in order to group the input data.

#### *b: RANDOM FOREST*

Random Forests is a classification and regression method based on the ensemble of a proliferation of decision trees [48]. Recently, attention has been given to ensemble learning, a method which can create several classifiers and produce aggregate results [49]. The two commonly known and used methods for the classification of trees are boosting as proposed by [50] and bagging as proposed by [51]. The latter proposed the general operating mechanism of the Random Forest (RF), which augments the additional layer of randomness to bagging. In RF, each tree is a standard Classification or Regression Tree (CART) that uses the so-called splitting criteria like Decrease of Gini Impurity (DGI). Moreover, it picks the splitting predictor from a randomly chosen subset of predictors. Each tree is fabricated by using a bootstrap sample of the data, and the prediction of all trees are ultimately accumulated by means of majority voting [48].

#### c: SUPPORT VECTOR MACHINE

Support vector machines are considered to be universal learners. Generically they learn linear threshold functions. However, with the use of a suitable kernel function plug in, they can be used to learn in different applications in the form of radical basic function and sigmoid neural networks and be trained on polynomial classifiers [52]. SVMs were initially intended for binary classification, but research has extended it into a multiclass classification. There are two commonly used approaches for multiclass SVMs. The first deals with fabricating and conjoining the number of binary classifiers, and the second one directly involves keeping all data in one optimization construction [53].

#### d: GENETIC ALGORITHM

The idea of developing the Genetic Algorithm was initiated and developed by John Holland. He proposed this idea in the year 1975 in his book ''Adaptation in natural and artificial systems''. Since then, the Genetic Algorithm (GA) has been increasingly recognized as a popular evolutionary computational research technique [54]. It has gained popularity over the years as an optimization tool in a variety of research domains, including computer science, operational research, engineering, management, and social sciences [55]. A major reason behind the realization of this technique is its diversified applicability, efficiency of operations, and applicability on global scenario [56].

Genetic Algorithms are search and optimization techniques in a multifaceted search space. They are inspired by the concept of genetics and natural selection. Some essential and principal ideas are adopted from the field of genetics and then used artificially to create a kind of algorithm which is flexible, robust, and efficient in nature [57]. Moreover, it characterizes the emergent technique which is to be used to understand different relationships in the course of the development of data, in which the data can come in the form of binary strings, formula, program, query, grammar, and images [55].

#### e: NAÏVE BAYES

The Naïve Bayes (NB) classifier has long been used in most applications of supervised machine learning. It is considered a tool for the retrieval of data [58]. It is based on a simple theorem of probability for making a probabilistic model of data. The mechanics of the NB algorithm are applied to numeric data [59]. It is simple, easy to understand, and quick for classification. It normally entails a minimal data set for training and then is used to predict the parameters needed for classification purposes.

#### f: DECISION TREE

The decision tree (DT) classifier has been widely used for prediction and classification of tasks. The rules in creating the decision tree are easy to understand. The classifiers built through the decision tree are given in hierarchical representation. The tree is composed of decision nodes, event nodes, edge, and path [60]. A variety of classifiers are used in a variety of applications. Some of the DT classifiers are ID3, C4.5, and C5.0, but a common problem that has been found in the DT classifier is the ability to incorporate all types of variations in data, including noise when trees get bigger and deeper. This problem is commonly known as overfitting. In addition, the structure of the tree is distorted with the addition of data. To avoid this problem, the random forest technique is recommended in which many trees are formed and trained by dividing the training sets, and outcomes are generated through aggregation of all trees.

#### g: K-NEAREST NEIGHBOUR

The K–nearest neighbor algorithm performs classification based on instance learning. It works through a non-parametric procedure of storing all inputs and instances and classifies new inputs by means of similarity measures like Euclidean distance [54], [55].

#### h: ENSEMBLE VOTED CLASSIFIER

An ensemble of classifiers is the collection of many classifiers in which their decisions are aggregated by means of weighted an unweighted voting mechanism to predict the outcome. The classification in the ensemble voted classifier is based on a voting mechanism in which classification of new instances is done by considering the majority vote of the prediction [51], [56]–[58].

### D. COMPARISON BETWEEN THE SENTIMENT ANALYSIS METHODS

Table 1 summarizes the various common methods used in the sentiment analysis literature.

The extant literature shows the effectiveness of the machine learning algorithm. However, with the increase in its effectiveness, the complexity of the method has also increased. Table 2 shows a comparison between the working principle, advantage, disadvantage, and time complexity of the most commonly used machine learning algorithms for sentiment analysis.

## VI. APPLICATION AREAS OF OPINION MINING AND SENTIMENT ANALYSIS

Opinion Mining and Sentiment Analysis has gained popularity in recent years and has been applied in many application areas. It has been used in diversified areas like healthcare, the financial sector, sports, politics, hospitality and tourism, and consumer behavior, as highlighted in Appendix 1. Some of the growing and emerging application areas will be reviewed in this section.

### A. HEALTHCARE

Korkontzelos *et al.* [70] enhanced the state-of-the-art adverse drug reaction method (ADR) by incorporating sentiment analysis features. The purpose is to examine the influence of sentiment analysis features in locating ADR mentions. For this, annotated ADR posts related to 81 drugs are collected from a Twitter and DailyStrength forum. Then, sentiment

**TABLE 1.** Summary of the common methods used in sentiment analysis.

| Studies | Machine learning-based classification | | | | | | | | Lexicon-based classification |
|---|---|---|---|---|---|---|---|---|---|
| | SVM | NB | KNN | RF | DT | ANN | CNN | LDA | |
| [3] | - | - | - | - | - | - | - | - | ✓ |
| [4] | - | - | - | - | - | - | - | - | ✓ |
| [5] | - | - | - | - | - | - | - | ✓ | - |
| [6] | ✓ | - | - | - | - | - | - | - | - |
| [7] | - | ✓ | - | - | - | - | - | - | - |
| [8] | ✓ | - | - | - | - | - | - | - | - |
| [9] | - | - | - | - | - | - | - | - | ✓ |
| [10] | - | ✓ | - | - | - | - | - | - | - |
| [11] | ✓ | - | - | - | - | - | - | - | - |
| [12] | ✓ | - | - | - | - | - | - | - | - |
| [13] | ✓ | - | - | - | - | - | - | - | - |
| [14] | ✓ | - | - | - | - | - | - | - | - |
| [15] | ✓ | ✓ | - | - | ✓ | - | - | - | - |
| [16] | ✓ | - | - | - | - | - | - | - | - |
| [17] | ✓ | - | - | - | - | - | - | - | - |
| [18] | ✓ | ✓ | - | - | ✓ | - | - | - | - |
| [19] | - | - | - | - | - | - | - | - | ✓ |
| [20] | - | - | - | - | - | - | - | - | ✓ |
| [21] | - | - | - | - | - | - | - | - | ✓ |
| [22] | - | ✓ | - | - | - | - | - | - | - |

analysis features are added to evaluate their effectiveness in locating ADR mentions. The results show that sentiment analysis features slightly boost the performance of ADR mention in both tweets and health-related posts in the forum, which can be utilized as pharmacovigilance practice in the future.

Rodrigues *et al.* [71] studied the mood of cancer patients by using a sentiment analysis tool named SentiHealth-Cancer. The aim of the study is to uncover the emotional condition of cancer patients though their discussion in the Portuguese language among the Brazilian online community. The posts are collected from the two different cancer communities on

**TABLE 2.** Comparison of working principle, advantage, disadvantage, and time complexity of most commonly used machine learning algorithms for sentiment analysis.

| *Machine learning Algorithms* | Working principle | Advantage | Disadvantage | Time complexity |
|---|---|---|---|---|
| *ARTIFICIAL NEURAL NETWORK (ANN)* | The system in ANN encompasses closely knit adaptive processing elements, also called artificial neurons or nodes which are proficient in carryout enormous analogous and parallel computations for the purpose of information processing and knowledge representation [47]. | With developments in deep learning such as recurrent, feed forward, and CNN, ANNs are applied successfully in many cases. ANN and its family algorithms give better accuracy in many sentiment applications, providing that they have enough data to train them. | ANNs entail more processing power and are mainly employed on graphics processing units. | $O(emnk)$ where $n$ shows number of instances, $m$ specifies number of features, $e$ denotes number of epochs and $k$ represent number of neurons [66], [67]. |
| *RANDOM FOREST* | Random Forests is a classification and regression method based on the ensemble of a proliferation of decision trees [48]. | Effective defense against over-fitting; ability to use many features, therefore no need to apply feature selection algorithms; the variance of the method reduces as the number of trees in the forest grows, while the bias is still the same [67]. | Low model interpretability, because of correlated variables performance may be reduced [67]. | $O(Tmnlogn)$ where $n$ represents number of instances, $m$ represents number of features, $T$ represents number of trees [43]. |
| *SUPPORT VECTOR MACHINE* | Generically it learns linear threshold function. However, with the use of suitable kernel function plug in, it can be used to learn in different application which are in the form of radical basic function, sigmoid neural networks and be trained on polynomial classifiers [52]. | SVM provides considerably fast classifiers; has potential generalization ability; is practically suitable for cases in which the number of features is higher than the number of instances. | A limitation of SVM is mainly related to the selection of kernel. | $O(n^2)$ where $n$ represents the number of instances [68]. |
| *GENETIC ALGORITHM* | Genetic Algorithms are search and optimization techniques in multifaceted search space. They are inspired by the concept of genetics and natural selection. That is, some essential and principle ideas are adopted from the field of genetics and then used artificially to create a kind of algorithm which is flexible, robust, and efficient in nature [57]. | The major reason behind the realization of this technique is its diversified applicability, efficiency of operations, and applicability to global scenarios [56]. | No guarantee of finding global maxima, requires even more time convergence. | $O(gkmn)$ where $n$ represents the number of instances, $m$ represents the number of features, $g$ represents the number of generation, and $k$ represents the size of populations [67]. |
| *NAÏVE BAYES* | This method is based on simple theorem of probability for making a probabilistic model of data. The mechanics of NB algorithm are applied to numeric data [59]. | NB is a simple, easy to understand, and quick technique for classification. It normally entails minimal data set for training and then is to be used to predict the parameters needed for classification purposes. | The main disadvantage of using NB is its susceptibility to Bayesian poisoning, a technique used to decrease the effectiveness of a system which relies on NB algorithms. Therefore, spammers can have regenerated biased sentimental posts that are difficult to detect from the original sentiment. | $O(nm)$ where $n$ represents the number of instances and $m$ represents the number of features [43]. |
| *DECISION TREE* | The rules in creating the decision tree are easy to understand. The classifiers that are built through the decision tree are in hierarchical representation. The tree is composed of decision nodes, event nodes, edge, and path [60]. | Decision trees are easy to understand. | The common problem that has been found in the DT classifier is the ability to incorporate all types of variations in data, including noise when trees get bigger and deeper. This problem is commonly knowns as overfitting. In addition, the structure of the tree would be distorted with the addition of data. | $O(mn^2)$ where $n$ represents the number of instances and $m$ represents the number of features [67], [69]. |
| *ENSEMBLE VOTED CLASSIFIER* | An ensemble of classifiers is the collection of many classifiers in which their decisions are aggregated by means of weighted and unweighted voting mechanisms to predict the outcome. | The final predication is based on voting from many classifiers; therefore, the error reduces the learning capability during the training and is enhanced as different learning theorems are used. | It increases the complexity of the system compared to single classifiers. | Depends on the classifiers used to construct the model. |

Facebook. The proposed tool, SentiHealth-Cancer (SHC-pt), is compared with AlchemyAPI 1.1.4v, Seman-tria 3.0.67v, SentiStrength 0.1v, and Textalytics 1.2 v in six different experiment settings. The result shows that this tool outperformed other sentiment analysis in all experiments including the complex setting, which shows that this tool is far better than other tools, especially in the cancer context.

Kim *et al.* [72] examined the coverage and sentiment trends of different media sources. The study relied on two sources: news publications and Twitter. The aim of the study is to evaluate the difference in coverage on issues like the Ebola virus and investigate the difference in sentiments on both mediums and to check the degree of change of sentiments over different time periods. For the experiment, about 16,189 news articles from around 1006 different publication sources and 7,106,297 tweets were collected. The experiment revealed that both Twitter and news media are different mediums of communication in terms of coverage and sentiment dynamics. The experiment on Twitter indicated that it is a time-bound medium and seems to be narrower in terms of coverage and have a shorter life span than news media.

### B. FINANCIAL SECTOR

Hai *et al.* [73] developed a stock price prediction model by incorporating sentiments of specific topics related to the company. Two datasets, including historical price dataset and mood information dataset, have been used to evaluate the effectiveness of the model. Historical stock prices of 18 companies extracted from Yahoo Finance and messages related to stock prediction, mood of the investors, discussion related to management of the organization, and specific events were extracted from the Yahoo Finance message board. The support vector machine (SVM) was used as a classifier, and six features were applied, which include price, human sentiment, sentiment classification, Latent Dirichlet Allocation (LDA) based method, joint sentiment/topic (JST) based method and aspect-based sentiment were designed and incorporated to evaluate the effectiveness of sentiment analysis in predicting stock market movement. The result shows that the performance accuracy of the proposed sentiment drove stock price model is approximately 2 percent better than the model, which only uses historical prices. Moreover, accuracy in predicting difficult stocks is approximately 10 percent better than historical price driven method.

Li and Meesad [74] also worked on predicting stock market trends, but the study focused on reducing socialization biasness by proposing a model which blends both sentiment analysis and socialization biasness through inverse bias algorithm. The proposed model is based on the support vector machine with hybrid features. For the experiment, 4,622 tweets were collected from Topsy.com during the timeframe from July 1, 2013 to May 30, 2014. The result indicated that SVM functioned better than other algorithms like Naïve Bayes and K-nn. Moreover, the model based on linear SVM classification with hybrid feature selection has improved its accuracy (from 84.06% to 86.95%) in predicting stock market movement when using 10 – folder cross validation. In contrast the accuracy of inverse socialization bias also increased from 86.95% to 90.33% while incorporating with linear SVM algorithm.

Chen *et al.* [75] investigated the relationship between public opinions or emotions with the changes in the stock market of China. The experimental dataset used for this study is Weibo, a social media website. Initially, the Chinese segmentation tool Jieba, which is based on dynamic programming, is used to fragment the text into different words and then categorize them into seven basic emotion categories which is based on Chinese Emotion Word Ontology. The result based on Granger causality analysis suggested that emotional states like happiness and disgust are strongly predictive to stock market prices in China. Moreover, the study demonstrated that, with the use of basic neural network model, emotional states aids to predict movement in the stock prices.

### C. SPORTS

Schumaker *et al.* [76] studied on prediction of English Premier League matches outcomes through a central sport system based on the crowdsourced sentiment technique. Moreover, the system also used it for wagering decision purposes. For the experiment, twenty clubs were selected, and the tweets from the last three months of the English Premier League season were analyzed. During that time, 122 matches were played, and about 96 hours of tweets containing the club hashtag before each match were collected from the Twitter API. During that time period, 18,027,966 tweets were collected. The results suggested that crowdsourced sentiment outperformed in predicting the match outcome than the crowdsourced odds. The analysis suggests that tweet sentiment outshines odd favorite wagering and has better payout returns in comparison to non-favorite wagering. Furthermore, goal difference and net payout return are improved with overwhelming positive sentiment. The analysis of the study proposes that professional odds markedly forecast non-positive match outcomes and show close goal margins.

Yu and Wang [77] studied sentiment analysis on tweets of US sports fans during five games of the FIFA World Cup 2014. The objective of the study was to evaluate their emotional responses and changes during the games, especially after their own or opponents' goals. For this purpose, several tweets were collected from three US games and two non-US games for comparative analysis. The results suggested that fear and anger were the dominant emotional state that US fans expressed when the opposing team scored, and that emotional state decreased with the US team goal result. Moreover, emotional responses of US fans in non-US games were ambiguous, but surprisingly, fans' tweets are more inclined towards positive emotions like joy and anticipation than negative emotions. The result of the study showed consistency with predictions of disposition theory, which states that emotions can be more predictable with the clarity of the fan base.

### D. POLITICS

Ceron *et al.* [78] used social media forums to investigate and follow the political preferences of the general public. The study focused on examining the pattern of the online popularity of different Italian political leaders throughout the year of 2011 and analyzed the intention of French voters in choosing the candidate for French Presidential ballot in 2012 and then following legislative elections. The study employed Hopkins and King's method of analyzing tweets. The result showed that social media platforms have a great capacity to predict electoral outcomes, and the result is highly correlated with old-fashioned methods like mass surveys. In addition, this study also emphasized and elucidated the increasing power of predictability through social media data analytics with the increase of opinions and expressions online.

Alfaro *et al.* [79] worked on opinion mining and sentiment analysis by integrating supervised and unsupervised machine learning techniques through a multi-stage procedure for the purpose of automatic recognition of various opinion trends in the weblogs. The study employed faculty and university weblog comments to test the framework. The test result indicated that SVM classification technique outperformed k-NN in terms of accuracy; however, integrating both approaches may increase accuracy. Based on the result, Alfaro *et al.* [79] anticipated that the proposed technique and procedure may be applied in different domains from the electoral campaign to develop a public policy or new law. Furthermore, it can be applied in analyzing feedback on a company's products and services and can be linked to the company market campaigning activities.

### E. HOSPITALITY AND TOURISM

Philander and Zhong [80] studied sentiment analysis on Las Vegas resorts as a case study. The study employed Twitter data to reveal the application of sentiment analysis as low cost and a real-time tool in evaluating customer perceptions about the services. Using Twitter data, a sentiment index was created using a sentiment lexicon methodology. The resulting sentiment metrics were used for performance comparative analysis of different firms over different time periods. The outcome of the sentiment score is then compared with data from TripAdvisor to evaluate external validity, which shows that both sentiment metrics and TripAdvisor are very much the same in terms of convergent and discriminant validity. The analysis shows that Twitter contains broad, direct, and indirect perspective about people's opinions towards Las Vegas properties, unlike TripAdvisor, which mainly focuses on the hotel customers' experiences and perception of facilities and services.

Rodolfo *et al.* [81] utilized text comments with the aim of applying sentiment analysis through three different algorithms in predicting overall hotel rankings. For this purpose, the numerical ratings and more than a million reviews of hotels situated in seven cities were gathered from the TripAdvsior web portal. Reviews were deemed as positives and negatives by using three different sentiment analysis tools. The results indicated that all classifiers are positively correlated with the actual rating of TripAdvisor to support the argument that the textual data of users can be transformed into numerical ratings. Moreover, the reliability of all three classifiers in predicting the hotel rating were compared with actual ratings and found complex algorithms, which, based on boosting and recursive neural tensor networks, were better in performance in relation to simple Naïve Bayesian algorithm. The study would be useful in areas like traveller forums by simply summarizing and consolidating opinions of potential and existing customers and can be used for predicting the rating.

### F. OTHER APPLICATION AREAS

Chung and Zeng [82] developed a framework for policy makers and evaluated public opinions about US immigration and border security on social media through a system called iMood based on sentiment and network analysis. For the experiment, around 909,0350 tweets were evaluated on the grounds of 300,000 users' sentiments and emotions in three different phases starting from May 2013 and ending in November 2013. Based on the analysis, the study highlighted significant changes in emotion and sentiment among different time zones or phases. The study would be a starting point in framing a policy in relation to public sentiments.

Liang *et al.* [83] employed a multifaceted sentiment analysis in predicting the sales of mobile apps based on online word-of-mouth textual comments. The comments related to product quality and service quality of 79 paid and 70 free apps are extracted from the iOS store. The test indicated that product and service quality are the dominant features in the eyes of the user and have a significantly positive impact on the ranking of app sales.

Kim *et al.* [84] mined public opinion from a social media networking platform in predicting box office performance based on a movie trailer. For the experiment, comments or reviews of trailers for 200 different movies were gathered from YouTube to analyze sentiments about movies. Moreover, the marketing property data of the chosen movies were collected from IMDB and the box office website, and these movies were categorized into different classes based on business. The results suggested that combining viewer comments and marketing properties assisted in improving prediction about box office performance.

Zhou *et al.* [85] exploited multi-granularity sentiment analysis approach in understanding the behaviors of consumers in different regions. The study conducted on Chinese and American customers of digital camera, smart phones, and tablet computers. The study revealed interesting findings, including that American customers are blunt and generally direct in expressing their views about the product features. On the other hand, Chinese customers most of the time used soft expressions. Moreover, American customers emphasized product details and internal features more, while the Chinese paid more attention to general feelings and the external features of products.

D'Avanzo and Pilato incorporated a cognitive approach based on collaborative learning to improve buyers' online shopping decisions. The proposed approach is experimented on Facebook reviews of two markets: the smart phone brand Nokia and the fashion brand Zalando and Privalia. The study employed Bayesian social sentiment technique and summarized opinions from different markets in order to arrive at a specific decision. This would help speed up the shopping activity and ultimately enhance the online shopping decision. The authors claimed that manufacturers or sellers would benefit through this approach, either by improving their offerings to the buyers or changing the product lines based on customers' feedback.

## VII. OPEN CHALLENGES

In this section we discuss the open challenge of big data sentiment analysis. We discuss both technical (challenges related to the development of Sentiment technique) and non-technical challenges (challenges related to the application of Sentiment analysis).

### A. CHALLENGES RELATED TO THE DEVELOPMENT OF SENTIMENT TECHNIQUE

In this section, we discussed the challenges related to the development of sentiment and opinion classifiers. The following section will discuss first the heterogeneous characteristics of big data-related challenges and the future directions needed to develop a method that can effectively deal with the heterogeneous characteristics of big data. The second challenge related to users' network (future researches need to further investigation on how to make use of user' relationship network for enriching the output of the sentiment system). Third challenge related to analyzing sparse, uncertain, and incomplete data and future researches are suggested to focus on developing sentient analysis method which can effectively be able to handle sparse, uncertain, and incomplete data. Lastly Semantic relations in multi-source data fusion.

#### 1) HETEROGENEOUS CHARACTERISTICS OF BIG DATA

Big Data occupies several characteristics but massive amount of data in heterogeneous and diverse formats (unstructured data) are the notable characteristics of big data. Such characteristic is created due to diverse strategies of how the data is collected, from where the data is collected and nature of diverse applications. Considering this characteristics, the sentiment classifier should be work effectively with such heterogeneous characteristic of big data. Usually traditional sentiment classifier was dealing with data from one source for example the company online review or company feedback records. However, in the era of big data, the sentiment classifiers should have the ability to handle diverse data from different data sources.

Big users' network,

In most above discuss studies have analyzed the post in order to classify the sentiment of the post regardless how an opinion is formed. However, with introduction of with big data such as big social data from social media website such twitter, rich information of people and their network with such platform can drive the researchers into deeper analysis of how the sentiment is formed. Individuals form similar friend clusters constructed according to on their common hobbies, views, interests, family relation or geographical relation. Such social contacts usually occur in not only our everyday activities, but similarly remain very common in virtual worlds [86]. The impact of users networks in forming the sentiments need more Investigation which may hold significance important strategies for smarter marketing method which does not only understand the opinion polarity of a post but also understands how this post is constructed within the social connections.

#### 2) ANALYZING SPARSE, UNCERTAIN, AND INCOMPLETE DATA

One of the characteristic of big data is that it contains a lot of noise because of the wide use of abbreviations and misspellings. This phenomenon, known as data sparsity [40], has an effect on the accuracy of the sentiment classification. Moreover, the big data (for example social media sites) are incomplete due to the privacy restriction. Therefore, some information such as location that can be used as features to enhance the functionality of sentiment classifier. For example, building sentiment analysis of political posts within specific country needs to understand the location of the posts. However, geolocation information may exist in some posts and not in other posts. Therefore, future works are required to create sentiment classifiers which should be able not only to classify the sentiment polarity of a message but also to predict the incomplete information to provide more accurate details to construct a complete application at the end.

#### 3) SEMANTIC RELATIONS IN MULTI-SOURCE DATA FUSION

Twitter, Facebook, Instagram, and YouTube may discuss an event at the same time. Analyzing semantic relations in these data sources can offer better insights and better understanding of whole sentiment picture. Analysis of an event from different sources [86] and then constructing semantic associations between text, image, and video data will significantly improve and enrich the output of the sentiment analysis systems [86]. Nevertheless, it is a challenge to construct such a semantic association based on models to fill the semantic gap between such heterogeneous data sources [86]. This can introduce a great opportunity to create sentiment analysis models based on multi-source data fusion.

### B. CHALLENGES RELATED TO THE APPLICATION OF SENTIMENT ANALYSIS

#### 1) DOES SENTIMENT ANALYSIS HELP IN DESIGNING ENTERPRISE STRATEGIES?

Big data has become an important element for enterprises to understand customers' opinions about their products. Big data sources such as social media provides huge user-generated data which is a worthy source of opinions and are treasured for many applications that involve comprehension

**Table 3. Synthesized areas of opinion mining and sentiment analysis applications.**

| Category | S. No | Source | Purpose | Dataset(s) | Area |
|---|---|---|---|---|---|
| **Health Care** | **1** | [70] | To explore the outcome of sentiment analysis features in locating ADR mentions. | Daily Strength and Twitter | Drugs |
| | **2** | [71] | To detect the mood of cancer patients in online communities | Facebook | Medical Science |
| | **3** | [72] | To investigate the topic coverage and sentiment dynamics of two different media sources on Health issue like Ebola | Twitter and Factiva Press Release Service (Online) database for news articles | Health care |
| **Politics** | **4** | [79] | To use sentiment analysis and opinion mining on weblog comments by linking supervised and unsupervised machine learning approaches | Faculty weblogs | Policy making and Electoral campaign |
| | **5** | [78] | To analyze citizens' political preferences using Twitter data | Twitter | Political preference of general public |
| **Financial Sector** | **6** | [74] | To predict stock market trends by linking sentiment analysis with socialization bias in social networks | Tweets collected from Topsy.com | Stock market |
| | **7** | [75] | To forecast the stock market price in China by using public emotion or opinions on Chinese social media websites | Chinese website weibo.com and http://finance.sina.com.cn/ | Stock market |
| | **8** | [73] | To develop a model for predicting the movement of stock price by means of social media sentiment data. | Yahoo Finance message board | Stock market |
| **Hospitality and Tourism** | **9** | [80] | To develop efficient and real-time measures of hospitality customer attitudes or perceptions about the resorts using Twitter sentiment analysis | Twitter | Hospitality and tourism |
| | **10** | [81] | To evaluate the reliability of numerical assessments of hotels computed through sentiment analysis algorithms | Tripadvisor.com | Hospitality |
| | **11** | [95] | To propose an aspect-based opinion mining approach in tourism product reviews | Trip advisor | Tourism |
| **Sports** | **12** | [76] | To analyze sentiments in the tweets and predict match outcomes in English Premier League matches | Twitter | Sports |
| | **13** | [77] | To examine U.S. soccer fans' emotional responses in their tweets during the course of a match. | Twitter | Sports |
| **Identifying, Locating and Forecasting** | **14** | [82] | To analyze twitter sentiments of opinion leaders, influential users and community activists about US immigration and border security policy | Twitter | Border security |
| | **15** | [96] | To exhibit a Social Sentiment Sensor (SSS) system on the application of Sina Weibo for detection of everyday hot topics and then investigate the distribution of sentiment toward these topics | Sina Weibo | Detection of hot topics and their distribution |
| | **16** | [97] | To examine hotspot detection on online forums and develop a forecast by means of sentiment analysis and text-mining approaches | Online Sina Sports community | Locating and forecasting |
| | **17** | [84] | to propose a scheme of opinion mining for predicting the box office performance of newly launched movie from online comments on the movie's trailer | YouTube and IMDB website | Predicting movie performance |

**Table 3.** *(Continued.)* **Synthesized areas of opinion mining and sentiment analysis applications.**

| | | | | | |
|---|---|---|---|---|---|
| | **18** | [98] | To predict client entities' ability to continue as a growing concern through data mining | Worldscope database - companies listed in AMEX, NASDAQ and NYSE | Prediction |
| **Reviews** | **19** | [85] | To investigate consumer behavior using opinion mining techniques | Amazon | Product reviews |
| | **20** | [99] | To extract explicit and implicit feature opinions from reviews | Trip advisor | Reviews related to travelling |
| | **21** | [100] | To analyze user-generated multi lingual opinions through opinion mining approach | YouTube | Reviews |
| | **22** | [101] | To propose a framework for Cantonese reviews | Hong Kong local forums | Reviews |
| | **23** | [102] | To propose the opinion-mining method based on feature-based sentiment classification in extracting the online electronic word-of-mouth on weblogs in Taiwan | Yahoo blog and wretch | Reviews on hotel |
| | **24** | [103] | To study the effectiveness of the classification of Spanish opinion in five different categories | Corpora | Movie and technological product reviews |
| | **25** | [104] | To conduct sentiment analysis using EOSenti Miner, an ontology is applied based on Chinese online reviews from a semantic perspective | www.dianping.com<br>www.JD.com | Online reviews |
| | **26** | [105] | To propose an NLP-based syntactic approach for opinion mining on Spanish reviews | Corpus | Spanish reviews |
| | **27** | [106] | To propose and develop an efficient sentiment analysis model that has the ability to extract product aspects and investigate sentiments from these aspects | Canon, Creative, Apex, Nikon, and Nokia | Customer reviews about the product |
| **Marketing and Sales** | **28** | [107] | To analyze not only the user-generated reviews about product recommendation but also expressions of sentiment that are associated with product features | Amazon | Product recommendations |
| | **29** | [108] | To propose an opinion mining extraction algorithm and investigate the attributes among heterogeneous products in the same category. | Amazon | Analysis of product features |
| | **30** | [109] | To propose a novel model in gauging the resemblance between online movies and TV episodes by means of microblog | Netflix Prize | Predicting viewing behavior |
| | **31** | [110] | To mine users opinions in helping buyer's shopping decision | Facebook page of Nokia, Zalando, and Privalia | Buying behavior |
| | **32** | [111] | To propose an expert recommendation system for user-generated product reviews from opinion extraction to suggestion and comparing reviewed products to recommending the product. | epinion.com<br>Data of printers, users, and reviews | Recommendation |
| | **33** | [83] | To inspect how the sentiments of various topics posted in online reviews affect sales of mobile app | Collected a data set on iOS app sales in Taiwan from App Annie (www.appannie.com) | Predicting sales |
| | **34** | [112] | To propose a method of estimating connection between user profiles, their value structures, and attitudes based on their replies and comments | YouTube | Reviewing attitude - Application in marketing |
| **Assessment and Evaluation** | **35** | [113] | To implement sentiment analysis of Thai children stories using SVM | Thai stories dataset | Assessment of techniques in different datasets |
| | **36** | [114] | | Twitter, Digg and Myspace | Assessment of techniques in different datasets |

**Table 3.** *(Continued.)* Synthesized areas of opinion mining and sentiment analysis applications.

| | | | | | |
|---|---|---|---|---|---|
| | **37** | [115] | | Movie review dataset and Twitter dataset | Assessment of techniques in different dataset |
| | **38** | [116] | To evaluate the effectiveness of integrated sentiment analysis apporoach | Large movie review dataset; IMDB | Assessment of technique; apply to movie review dataset |
| | **39** | [117] | To evaluate the effectiveness and accuracy of eSAP in different datasets | Cornell Movie Review dataset, large movie review dataset, multi-domain sentiment datasets | Assessment of technique; apply to movie review dataset and multiple products |
| | **40** | [118] | Propose a visual sentiment topic model for microblog images | Sina Microblog | Assessment of technique |
| | **41** | [119] | To propose an approach for the detection of sentiments both at the entity and tweet levels. | Obama-McCain Debate, Health care Reform dataset (HCR) Standford Sentiment Gold Standard (STS-Gold) | Assessment of approach |
| | **42** | [120] | To capture public sentiments in the form of Arabic tweets by using the hybrid approach | Twitter API for Arabic tweets | Evaluating sentiments of Arabic tweets |
| | **43** | [84] | To analyze machine learning algorithms with opinion mining and sentiment analysis | Movie section of portal websites in Korea (http://nate.com and http://movie.daum.net) | Comparative analysis of techniques |
| | **44** | [121] | To integrate multiple classifiers in sentiment analysis | Apparels (A), books (B), DVDs (D), electronics (E), health appliances (H) and Sport-outdoor appliances (S). | Evaluate the effectiveness of integrating multiple classifiers |
| | **45** | [122] | To compute the pairwise word dependencies and develop an association-based unified framework that can recognize both implicit and explicit attributes to identify features from reviews. | Chinese product website and Chinese travel portal | Analysis of techniques |
| | **46** | [123] | To develop a social media analytics frameworks for opinion mining | Collected 14,204 items of social media content including blogs, forum (café) messages, and media news articles from January 2012 to June 2013 | Development of framework for social media analytics on Noodles product |
| | **47** | [124] | To study the effectiveness of different text pre-processing and OM algorithms for social media channels | Facebook, Twitter, blogs, discussion forums, and product review portals (Amazon and product review pages in 4 different languages) | Applicability of techniques on different social media channels |
| | **48** | [125] | To propose a unique method for the identification of features from online reviews through IEDR | Chinese cell phone and hotel forum websites | Assessment of technique |
| | **49** | [126] | To propose hybrid classification scheme for opinion mining | Twitter | Sentiment classification on politics, sports, and country in general |
| | **50** | [127] | To propose an intuitive, and unsupervised, lexicon-based approach that examines the degree of emotional intensity confined in text in order to make a prediction | Twitter, MySpace, Digg | Effectiveness of technique |
| **Multiple Applications** | **51** | | To propose an innovative approach in predicting sentiments in online textual messages. | 1. The Cornell Movie Review dataset 2. Obama - McCain Debate dataset 3. The SemEval-2015 Task 10 dataset | Movie/Election debate |
| | **52** | [128] | To develop an opinion formation framework an apply on three different scenarios | Blogs | Political discussion in Poland, Governance of Java standard, BP oil spill |
| **Government** | **53** | [129] | To determine whether citizens' sentiment can influence their participation with government through social media | Twitter | Citizens' involvement in US local government |

**Table 3.** ***(Continued.)*** **Synthesized areas of opinion mining and sentiment analysis applications.**

| | | | | | |
|---|---|---|---|---|---|
| **Construction Project** | **54** | [130] | To analyze the perspectives of public opinions related to hydro project | Microblogging website weibo.com | Construction Project |
| **Festival** | **55** | [131] | | Twitter and survey response | Literature Festival/Cultural Event |
| **Media** | **56** | [132] | To mine opinion polarity of general public comments in e-newspaper | 815 Arabic comments from local newspaper | Media |
| **Miscellaneous** | **57** | [133] | To develop super network model for opinion mining and analyze its application on super edge prediction. | Sina Weibo microblogs | Murder or suicide/Public opinion |
| | **58** | [134] | To analyze the positive or negative influence of popular or influential twitter user to their audiences. | Twitter API | Popular Twitter users' influence |

of the public opinion about events, products, persons, etc. For example, an enterprise that may capture the opinions of clients about their products [40] may apply some sentiment analysis to advance the quality of their products. However, more investigation is required to understand more factors that can work together with sentiment factors to better understanding the public opinion. For example, product "A" has many positive sentiments in the recent three months, but does it mean that the enterprises which produced product "A" need to maintain the quality of this product, as it still has a positive sentiment by the users, and does the enterprise need to increase the production of this product to meet the popularity of the product? Sentiment analysis alone may not be enough to answer these questions. Therefore, factors like competitors' information, country economic growth, consumer confidence index, and other factors may influence conclusive decisions. However, future studies need to integrate sentiment analysis methods with other factors to create effective methods for a fully automated decision-making system based on several factors. Moreover, future research needs to investigate whether big data-based methods can replace traditional methods (questionnaire-based methods) in making decisions or whether both can work best when integrated into a system.

#### 2) THE INFLUENCE OF THE POST

A user (A) with a large number of followers posts a negative comment about product X, and a user (B) with few followers posts a positive comment about the same product. Statistically, we have same number of positive and negative comments of product X, but in reality, the negative post should have more impact, since user A has several followers who can be influenced by his/her opinion compared to user B, who has few followers; consequently, his opinion influences fewer users compared to user A. Moreover, in many cases, the direct number of followers does not indicate high influence [83], [84]. Therefore, future research should further investigate the influence of sentiment polarity within the connected networks of users. Integrated sentiment analysis and influence measurement can provide a precise sentiment measurement which takes into consideration both the polarity of the post as well as its influence to capture consumers' views on a larger and deeper scale.

#### 3) THE IMPACT OF SOCIAL BOTS

Recently social bots have become sophisticated as well as threatening. The presence of social bots, especially in social media, can be a threat to online systems as well as to our society [89]. Social bots are software programs aimed to simulate human user on social media websites. Gradually, politicians, marketer, and different firms use these automatic social bots in online environments in order to manipulate public opinion [90]. For example, political bot-based strategies are used to enormously increase politicians' followers on social media sites and generate positive comments to create impressions of popularity [90]. Similarly, this method can be applied in other areas as well. Therefore, two questions may arise: first, is the output of sentiment analysis based on the genuine opinion of the user or is it a bot-generated opinion? Second, to which extent can the social bot influence public opinion? Hence, in future research, these two questions must be investigated in order to know how the opinion is formed and to recognize how sentiment analysis should be designed to take this factor into consideration.

## VIII. DISCUSSION AND CONCLUSION

A few conclusions can be made based on the findings of the systematic literature have been conducted on articles published in the Web of Science from 2000-2016 on OMSA. In addition, more articles have been published on sentiment analysis as compared to opinion mining since 2015. The emergence of the significance of sentiment analysis matches the development of social media usage, including reviews, forum blogs, micro-blogs, Facebook, Twitter, and other social networks. Interestingly, presently we have access to a huge amount of opinionated data which can further be used for different methods of analysis. Typically more than 80% of social media data can be monitored for analysis purposes [135].

For instance, a tweet contains a maximum of 140 characters, therefore monitoring software can assign a specific sentiment score to that tweet. That sentiment score represents a semantic judgment to examine whether the tweet seems to be positive, negative, or neutral. Gathering reviews on products and services on the other hand is mainly done using opinion mining. Products and services are considered as entities, and the process of mining opinions usually involves performing opinions of the texts. The trends further indicate that society is more likely to express feelings rather than what they think on social media platforms.

In terms of applications, more research has been done on the assessment or evaluation of the various methods of opinion mining and sentiment analysis. Although these refer to the evaluation of the techniques used, the data sets extracted from users' application databases thus include an element of human application. Marketing-related activities still dominate the applications followed by the financial, healthcare, and hospitality and tourism industries. It is further noted that applications of opinion mining and sentiment analysis for politics and government views are still growing. The data generated by people through their views on current events in the country can be very useful to political parties and for the general public in resurrecting the future course of action for their interests.

In the datasets used for OMSA, data from Twitter seemed to dominate the data sets. This is further aligned with the growth of the sentiment analysis in which most of the data is captured from social media more profoundly on Twitter as compared to the other data sources. Previously, Twitter has been used as a tool for disseminating and propagating information rather than simply a social networking site. Previous research shows that top users, as measured by the number of followers on Twitter, are mostly celebrities and those who attract the keen interest of the mass media [91]. An electronic platform for word-of-mouth influence in marketing [92], Twitter also serves as a political sentiment analysis predictor of elections [93] and as a stock market movement predictor [94]. This has widened its exploitation to politicians and other pundits. It can be used to quickly share information with people, promote new products, and communicate with celebrities' fans or political supporters.

## APPENDIX 1

See Table 3.

## REFERENCES

[1] R. Addo-tenkorang and P. T. Helo, "Big data applications in operations/supply-chain management: A literature review," *Comput. Ind. Eng.*, vol. 101, pp. 528–543, Nov. 2016.

[2] A. Zaslavsky, C. Perera, and D. Georgakopoulos. (2013). "Sensing as a service and big data." [Online]. Available: https://arxiv.org/abs/1301.0159

[3] J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Seoul, South Korea, Tech. Rep., 2011.

[4] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, "Vision and challenges for realising the Internet of Things," *Cluster Eur. Res. Projects Internet Things, Eur. Commision*, vol. 3, no. 3, pp. 34–36, Mar. 2010.

[5] A. Yasin, Y. Ben-Asner, and A. Menaeison, "Deep-dive analysis or the data analytics workload in cloudsuite," in *Proc. IEEE Int. Symp. Workload Characterization (IISWC)*, Oct. 2014, pp. 202–211.

[6] R. L. Villars, C. W. Olofson, and M. Eastwood, "Big data: What it is and why you should care," IDC, Framingham, MA, USA, White Paper 228827, Jun. 2011.

[7] M. Khoso. (2016). *How Much Data is Produced Every Day?* [Online]. Available: http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/

[8] P. Zikopoulos and C. Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY, USA: McGraw-Hill, 2011.

[9] P. Russom, "Big data analytics," *TDWI Best Pract. Rep.*, vol. 19, no. 4, pp. 1–34, 2011.

[10] T. H. Davenport and J. Dyché. (2013). *Big Data in Big Companies*. [Online]. Available: http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/bigdata-bigcompanies-106461.pdf

[11] Accenture. (2014). *Big Data Analytics in Supply Chain: Hype or Here to Stay?* [Online]. Available: https://acnprod.accenture.com/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Dualpub_2/Accenture-Global-Operations-Megatrends-Study-Big-Data-Analytics.pdf

[12] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[13] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, Dec. 2012.

[14] A. Misra, A. Sharma, P. Gulia, and A. Bana, "Big data: Challenges and opportunities," *Int. J. Innov. Technol. Exploring Eng.*, vol. 4, no. 2, pp. 41–42, 2014.

[15] M. Batty *et al.*, "Smart cities of the future," *Eur. Phys. J. Special Topics*, vol. 214, no. 1, pp. 481–518, Nov. 2012.

[16] R. E. Bryant, R. H. Katz, and E. D. Lazowska, "Big-data computing: Creating revolutionary breakthroughs in commerce, science and society," Comput. Community Consoritum, Washington, DC, USA, Tech. Rep. Version 8, Dec. 2008.

[17] F. R. Lucini *et al.*, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *Int. J. Med. Inf.*, vol. 100, pp. 1–8, Apr. 2017.

[18] Z. Khan and T. Vorley, "Big data text analytics: An enabler of knowledge management," *J. Knowl. Manage.*, vol. 21, no. 1, pp. 18–34, 2017.

[19] T. T. Thet, J.-C. Na, and C. S. G. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *J. Inf. Sci.*, vol. 36, no. 6, pp. 823–848, 2010.

[20] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2003, pp. 129–136.

[21] R. Piryani, D. Madhavi, and V. K. Singh, "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015," *Inf. Process. Manage.*, vol. 53, no. 1, pp. 122–150, 2017.

[22] A. Qazi, A. Tamjidyamcholo, R. G. Raj, G. Hardaker, and C. Standing, "Assessing consumers' satisfaction and expectations through online opinions: Expectation and disconfirmation approach," *Comput. Hum. Behav.*, vol. 75, pp. 450–460, Oct. 2017.

[23] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2006, pp. 244–251.

[24] A. Qazi, R. G. Raj, M. Tahir, E. Cambria, and K. B. S. Syed, "Enhancing business intelligence by means of suggestive reviews," *Sci. World J.*, vol. 2014, Jun. 2014, Art. no. 879323.

[25] M. Quigley, *Encyclopedia of Information Ethics and Security*. Hershey, PA, USA: IGI Global, 2007.

[26] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

[27] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.

[28] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[29] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, Sep. 2016.

[30] I. Chaturvedi, E. Cambria, S. Poria, and R. Bajpai, "Bayesian deep convolution belief networks for subjectivity detection," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 916–923.

[31] S. Poria, E. Cambria, D. Hazarika, and P. Vij. (2016). "A deeper look into sarcastic tweets using deep convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1610.08815

[32] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.

[33] D. Denyer and D. Tranfield, "Producing a systematic review," in *The Sage Handbook of Organizational Research Methods*. Thousand Oaks, CA, USA: Sage, 2009.

[34] D. Tranfield, D. Denyer, and P. Smart, "Towards a methodology for developing evidence-informed management knowledge by means of systematic review," *Brit. J. Manage.*, vol. 14, no. 3, pp. 207–222, 2003.

[35] R. J. Light and D. B. Pillemer, *Summing Up: The Science of Reviewing Research*. Cambridge, MA, USA: Harvard Univ. Press, 1984.

[36] D. M. Rousseau, J. Manning, and D. Denyer, "11 evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses," *Acad. Manage. Ann.*, vol. 2, no. 1, pp. 475–515, 2008.

[37] S. L. Newbert, "Empirical research on the resource-based view of the firm: An assessment and suggestions for future research," *Strategic Manage. J.*, vol. 28, no. 2, pp. 121–146, 2007.

[38] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar. 2013.

[39] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[40] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Comput. Surv.*, vol. 49, no. 2, p. 28, 2016.

[41] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining Text Data*. Boston, MA, USA: Springer, 2012, pp. 163–222.

[42] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[43] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[44] D. H. Wolpert and W. G. Macready, "No free lunch theorems for search," Santa Fe Institute, Santa Fe, NM, USA, Tech. Rep. SFI-TR-05- 010, 1995.

[45] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.

[46] C. A. L. Bailer-Jones, R. Gupta, and H. P. Singh, "An introduction to artificial neural networks," in *Automated Data Analysis in Astronomy*, vol. 52. New Delhi, India: Narosa Publishing House, Sep. 2001, p. 18.

[47] E. Grossi and M. Buscema, "Introduction to artificial neural networks," *Eur. J. Gastroenterol. Hepatol.*, vol. 19, pp. 1046–1054, Dec. 2007.

[48] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[49] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[50] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Stat.*, vol. 26, no. 5, pp. 1651–1686, 1998.

[51] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[52] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, vol. 98, 1998, pp. 137–142.

[53] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[54] S. N. Sivanandam and S. N. Deepa, "Genetic algorithms," in *Introduction to Genetic Algorithms*. New York, NY, USA: Springer-Verlag, 2007.

[55] M. Mitchell, "An introduction to genetic algorithms," *Sadhana*, vol. 24, nos. 4–5, pp. 293–315, Aug. 1999.

[56] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1989.

[57] A. Kumar, R. M. Pathak, and Y. P. Gupta, "Genetic algorithm based approach for file allocation on distributed systems," *Comput. Oper. Res.*, vol. 22, no. 1, pp. 41–54, 1995.

[58] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 4–15.

[59] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in *Proc. Learn. Text Categorization, Papers Workshop*, vol. 62, 1998, pp. 98–105.

[60] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[61] Y. Bao, N. Ishii, and X. Du, "Combining multiple k-nearest neighbor classifiers using different distance functions," in *Proc. IDEAL*, 2004, pp. 634–641.

[62] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, NY, USA: Academic, 2013.

[63] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 137–142.

[64] S. Danso, E. Atwell, and O. Johnson. (2014). "A comparative study of machine learning methods for verbal autopsy text classification." [Online]. Available: https://arxiv.org/abs/1402.4380

[65] W. L. Yeow, R. Mahmud, and R. G. Raj, "An application of case-based reasoning with machine learning for forensic autopsy," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3497–3505, 2014.

[66] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.

[67] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[68] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[69] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.

[70] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts," *J. Biomed. Inform.*, vol. 62, pp. 148–158, Aug. 2016.

[71] R. G. Rodrigues, R. M. das Dores, C. G. Camilo-Junior, and T. C. Rosa, "SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks," *Int. J. Med. Inform.*, vol. 85, no. 1, pp. 80–95, 2016.

[72] E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," *J. Inf. Sci.*, vol. 42, no. 6, pp. 763–781, 2016.

[73] T. Hai, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9603–9611, 2015.

[74] J. Li and P. Meesad, "Combining sentiment analysis with socialization bias in social networks for stock market trend prediction," *Int. J. Comput. Intell. Appl.*, vol. 15, no. 1, p. 1650003, 2016.

[75] W. Chen, Y. Cai, K. Lai, and H. Xie, "A topic-based sentiment analysis model to predict stock market price movement using Weibo mood," *Web Intell.*, vol. 14, no. 4, pp. 287–300, 2016.

[76] R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, Jr., "Predicting wins and spread in the Premier League using a sentiment analysis of Twitter," *Decis. Support Syst.*, vol. 88, pp. 76–84, Aug. 2016.

[77] Y. Yu and X. Wang, "World cup 2014 in the Twitter world: A big data analysis of sentiments in US sports fans' tweets," *Comput. Hum. Behav.*, vol. 48, pp. 392–400, Jul. 2015.

[78] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France," *New Media Soc.*, vol. 16, no. 2, pp. 340–358, 2014.

[79] C. Alfaro, J. Cano-Montero, J. Gómez, J. M. Moguerza, and F. Ortega, "A multi-stage method for content classification and opinion mining on weblog comments," *Ann. Oper. Res.*, vol. 236, no. 1, pp. 197–213, 2016.

[80] K. Philander and Y. Y. Zhong, "Twitter sentiment analysis: Capturing sentiment from integrated resort tweets," *Int. J. Hospitality Manage.*, vol. 55, pp. 16–24, May 2016.

[81] R. R. B. López, S. Sánchez-Alonso, and M. A. Sicilia-Urban, "Evaluating hotels rating prediction based on sentiment analysis services," *Aslib J. Inf. Manage.*, vol. 67, no. 4, pp. 392–407, 2015.

[82] W. Chung and D. Zeng, "Social-media-based public policy informatics: Sentiment and network analyses of US Immigration and border security," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 7, pp. 1588–1606, 2016.

[83] T.-P. Liang, X. Li, C.-T. Yang, and M. Wang, "What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach," *Int. J. Electron. Commerce*, vol. 20, no. 2, pp. 236–260, 2015.
[84] D. Kim, D. Kim, E. Hwang, and H.-G. Choi, "A user opinion and metadata mining scheme for predicting box office performance of movies in the social network environment," *New Rev. Hypermedia Multimedia*, vol. 19, nos. 3–4, pp. 259–272, 2013.
[85] Q. Zhou, R. Xia, and C. Zhang, "Online shopping behavior study based on multi-granularity opinion mining: China versus America," *Cogn. Comput.*, vol. 8, no. 4, pp. 587–602, 2016.
[86] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
[87] M. S. Khan, A. W. A. Wahab, T. Herawan, G. Mujtaba, S. Danjuma, and M. A. Al-Garadi, "Virtual community detection through the association between prime nodes in online social networks and its application to ranking algorithms," *IEEE Access*, vol. 4, pp. 9614–9624, 2016.
[88] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method," *Phys. A, Stat. Mech. Appl.*, vol. 468, pp. 278–288, Feb. 2017.
[89] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, 2016.
[90] S. C. Woolley, "Automating power: Social bot interference in global politics," *First Monday*, vol. 21, no. 4, Apr. 2016.
[91] S. Pei, L. Muchnik, J. S. Andrade, Jr., Z. Zheng, and H. A. Makse, "Searching for superspreaders of information in real-world social media," *Sci. Rep.*, vol. 4, Jul. 2014, Art. no. 5547.
[92] B. J. Jansen, M. Zhang, K. Sobe, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, Nov. 2009.
[93] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. ICWSM*, vol. 10, no. 1, 2010, pp. 178–185.
[94] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting stock market indicators through Twitter 'I hope it is not as bad as I fear,'" *Procedia-Social Behav. Sci.*, vol. 26, pp. 55–62, Jan. 2011.
[95] E. Marrese-Taylor, J. D. Velásquez, and F. Bravo-Marquez, "A novel deterministic approach for aspect-based opinion mining in tourism products reviews," *Expert Syst. Appl.*, vol. 41, no. 17, pp. 7764–7775, 2014.
[96] Y. Zhao, B. Qin, T. Liu, and D. Tang, "Social sentiment sensor: A visualization system for topic detection and topic sentiment analysis on microblog," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 8843–8860, 2014.
[97] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decis. Support Syst.*, vol. 48, no. 2, pp. 354–368, 2010.
[98] D. Martens, L. Bruynseels, B. Baesens, M. Willekens, and J. Vanthienen, "Predicting going concern opinion with data mining," *Decis. Support Syst.*, vol. 45, no. 4, pp. 765–777, 2008.
[99] F. Lazhar and T.-G. Yamina, "Mining explicit and implicit opinions from reviews," *Int. J. Data Mining, Modelling Manage.*, vol. 8, no. 1, pp. 75–92, 2016.
[100] A. Severyn, A. Moschitti, O. Uryupina, B. Plank, and K. Filippova, "Multi-lingual opinion mining on YouTube," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 46–60, 2016.
[101] J. Chen, D. P. Huang, S. Hu, Y. Liu, Y. Cai, and H. Min, "An opinion mining framework for Cantonese reviews," *J. Ambient Intell. Hum. Comput.*, vol. 6, no. 5, pp. 541–547, 2014.
[102] C. Chiu, N.-H. Chiu, R.-J. Sung, and P.-Y. Hsieh, "Opinion mining of hotel customer-generated contents in Chinese weblogs," *Current Issues Tourism*, vol. 18, no. 5, pp. 477–495, 2015.
[103] M. del Pilar Salas-Zàrate, E. López-López, R. Valencia-García, N. Aussenac-Gilles, Á. Almela, and G. Alor-Hernández, "A study on LIWC categories for opinion mining in Spanish reviews," *J. Inf. Sci.*, vol. 40, no. 6, pp. 749–760, 2014.
[104] W. Shi, H. Wang, and S. He, "EOSentiMiner: An opinion-aware system based on emotion ontology for sentiment analysis of Chinese online reviews," *J. Exp. Theor. Artif. Intell.*, vol. 27, pp. 423–448, Jul. 2015.
[105] D. Vilares, M. A. Alonso, and C. Gómez-rodríguez, "A syntactic approach for opinion mining on Spanish reviews," *Natural Lang. Eng.*, vol. 21, no. 1, pp. 139–163, 2015.
[106] A. Bagheri, M. Saraee, and F. De Jong, "Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews," *Knowl.-Based Syst.*, vol. 52, pp. 201–213, Nov. 2013.
[107] R. Dong, M. P. O'Mahony, M. Schaal, K. McCarthy, and B. Smyth, "Combining similarity and sentiment in opinion mining for product recommendation," *J. Intell. Inf. Syst.*, vol. 46, no. 2, pp. 285–312, 2016.
[108] H. Zhang, A. Sekhari, Y. Ouzrout, and A. Bouras, "Jointly identifying opinion mining elements and fuzzy measurement of opinion intensity to analyze product features," *Eng. Appl. Artif. Intell.*, vol. 47, pp. 122–139, Jan. 2016.
[109] H. Li, J. Cui, B. Shen, and J. Ma, "An intelligent movie recommendation system through group-level sentiment analysis in microblogs," *Neurocomputing*, vol. 210, pp. 164–173, Oct. 2016.
[110] E. D. Avanzo and G. Pilato, "Mining social network users opinions 'to aid buyers' shopping decisions," *Comput. Hum. Behav.*, vol. 51, pp. 1284–1294, Oct. 2015.
[111] A. Stavrianou and C. Brun, "Expert recommendations based on opinion mining of user-generated product reviews," *Comput. Intell.*, vol. 31, no. 1, pp. 165–183, 2015.
[112] H.-J. Jang, J. Sim, Y. Lee, and O. Kwon, "Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7492–7503, 2013.
[113] K. Pasupa, P. Netisopakul, and R. Lertsuksakda, "Sentiment analysis of Thai children stories," *Artif. Life Robot.*, vol. 21, no. 3, pp. 357–364, 2016.
[114] A. Muhammad, N. Wiratunga, and R. Lothian, "Contextual sentiment analysis for social media genres," *Knowl.-Based Syst.*, vol. 108, pp. 92–101, Sep. 2016.
[115] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level," *Knowl.-Based Syst.*, vol. 108, pp. 110–124, Sep. 2016.
[116] A. Ceron, L. Curini, and S. M. Iacus, "iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content," *Inf. Sci.*, vol. 367, pp. 105–124, Nov. 2016.
[117] F. H. Khan, U. Qamar, and S. Bashir, "eSAP: A decision support framework for enhanced sentiment analysis and polarity classification," *Inf. Sci.*, vol. 367, pp. 862–873, Nov. 2016.
[118] D. Cao, R. Ji, D. Lin, and S. Li, "Visual sentiment topic model based microblog image sentiment analysis," *Multimedia Tools Appl.*, vol. 75, no. 5, pp. 8955–8968, 2016.
[119] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 5–19, 2015.
[120] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis—A hybrid scheme," *J. Inf. Sci.*, vol. 42, no. 6, pp. 782–797, 2015.
[121] Y. Lin, X. Wang, Y. Li, and A. Zhou, "Integrating the optimal classifier set for sentiment analysis," *Soc. Netw. Anal. Mining*, vol. 5, no. 1, p. 50, 2015.
[122] Z. Hai, K. Chang, G. Cong, and C. C. Yang, "An association-based unified framework for mining features and opinion words," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 2, p. 26, 2015.
[123] Y. Kim and S. R. Jeong, "Opinion-mining methodology for social media analytics," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 1, pp. 391–406, 2015.
[124] G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, V. Stříteský, and A. Holzinger, "Computational approaches for mining user's opinions on the Web 2.0," *Inf. Process. Manage.*, vol. 50, no. 6, pp. 899–908, 2014.
[125] Z. Hai, K. Chang, J. J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 3, pp. 623–634, Mar. 2014.
[126] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," *Decis. Support Syst.*, vol. 57, pp. 245–257, Jan. 2014.
[127] G. Paltoglou and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, p. 66, 2012.
[128] P. Sobkowicz, M. Kaschesky, and G. Bouchard, "Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the Web," *Government Inf. Quart.*, vol. 29, no. 4, pp. 470–479, 2012.
[129] S. M. Zavattaro, P. E. French, and S. D. Mohanty, "A sentiment analysis of US local government tweets: The connection between tone and citizen involvement," *Government Inf. Quart.*, vol. 32, no. 3, pp. 333–341, 2015.
[130] H. Jiang, P. Lin, and M. Qiang, "Public-opinion sentiment analysis for large hydro projects," *J. Construct. Eng. Manage.*, vol. 142, no. 2, p. 05015013, 2015.
[131] B. Driscoll, "Sentiment analysis and the literary festival audience," *Continuum*, vol. 29, no. 6, pp. 861–873, 2015.

[132] A. M. Azmi and S. M. Alzanin, '''Aara'—A system for mining the polarity of Saudi public opinion through e-newspaper comments,'' *J. Inf. Sci.*, vol. 40, no. 3, pp. 398–410, 2014.
[133] Y. Liu, Q. Li, X. Tang, N. Ma, and R. Tian, ''Superedge prediction: What opinions will be mined based on an opinion supernetwork model?'' *Decis. Support Syst.*, vol. 64, pp. 118–129, Aug. 2014.
[134] Y. Bae and H. Lee, ''Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers,'' *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, pp. 2521–2535, Nov. 2012.
[135] Anderson. (May 30, 2017). *Sentiment Analysis or Text Analytics, What's the Difference Anyway?* [Online]. Available: http://nextgenmr.com/sentiment-analysis-text-analysis

**SHAHID SHAYAA** has accumulated over 13 years of experience working in the corporate sector in various business transformation projects in Accenture, Employees Provident Fund, Malaysia Airlines, and the Government of Malaysia's Performance Management Delivery Unit. He is currently the Founder and the CEO of Berkshire Media Sdn Bhd. He is instrumental in transforming the digital media industry through rigorous data driven approaches. He now advises the top leaderships on strategic communications for publicly traded companies and the public sector.

**NOOR ISMAWATI JAAFAR** is currently an Associate Professor and the Deputy Dean (Research and Development) with the Faculty of Business and Accountancy, University of Malaya. She has published papers in *Information & Management*, *Government Information Quarterly*, *Computers in Human Behavior*, *Telematics and Informatics*, *Behavior & Information Technology*, *Cyberpsychology, Behavior and Social Networking*, *Information Development*, and *International Journal of Mobile communications*. Her research interests include accounting information systems, information technology management, information technology governance, and social media.

**SHAMSHUL BAHRI** is currently a Senior Lecturer and the Head of the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. He is also a Prolific Qualitative Researcher and has published articles in high-ranked IS journals, such as *Information and Management* and *Information Systems Journal*. His areas of research include information, computer and communication technology and information systems (public health informatics).

**AININ SULAIMAN** is currently a Professor with the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. She has published papers in *Information and Management*, *Computers in Human Behavior*, *Telematics and Informatics*, *Behavior & Information Technology*, *American Journal of Scientific Research*, *Management Decisions*, *International Journal of Mobile Communications*, *Government Information Quarterly* and so on. Her research interest includes management information systems, technology diffusion, E-commerce, and green information technology.

**PHOONG SEUK WAI** is currently a Senior Lecturer with the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. She has published articles in the *International Journal of Advanced and Applied Sciences* and the *International Journal of Computing Science and Mathematics*. Her research areas are econometrics modeling, time series economics, and statistical modeling and spatial analysis.

**YEONG WAI CHUNG** is currently a Senior Lecturer with the Department of Operations and Management Information Systems, Faculty of Business and Accountancy, University of Malaya. He has published in many ISI-indexed journals, including *European Journal of Operational Research*, *Quality Technology and Quantitative Management*, *Computers & Industrial Engineering*, and the *Journal of Quality Technology* and *Quality Engineering*. His research areas are optimization, total quality management, operational effectiveness, and economic designs.

**ARSALAN ZAHID PIPRANI** received the B.Sc. degree (Hons.) in textile science from the Textile Institute of Pakistan in 2005, the M.B.A. degree with the majors in industrial management from the Institute of Business Management, Karachi, Pakistan, in 2009. He is currently pursuing the Ph.D. degree with the Department of Operations and MIS, University of Malaya, Malaysia, with the research area of supply chain management. He is also the Research Assistant with the University of Malaya for Social Media Data Analytics Project. His research interests include supply chain management, supply chain risk management, and social media data analytics.

**MOHAMMED ALI AL-GARADI** received the Ph.D. degree from the University of Malaya, Malaysia, in 2017. He has research experience of over four years. He has published over 10 research articles in refereed journals. His research interests include big data analytics, machine learning, deep learning, and the IoT and complex network analysis.