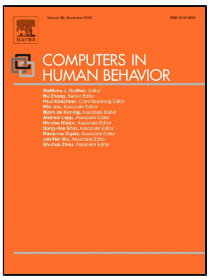
Accepted Manuscript



Big Social Media Data Analytics: A Survey

Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, Ejaz Ahmed

PII:	S0747-5632(18)30414-X
DOI:	10.1016/j.chb.2018.08.039
Reference:	CHB 5673
To appear in:	*Computers in Human Behavior*
Received Date:	12 December 2017
Accepted Date:	22 August 2018

Please cite this article as: Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, Big Social Media Data Analytics: A Survey, *Computers in Human Behavior* (2018), doi: 10.1016/j.chb.2018.08.039

Big Social Media Data Analytics: A Survey

Norjihan Abdul Ghani[1], Suraya Hamid[1], Ibrahim Abaker Targio Hashem[2], Ejaz Ahmed[3]

norjihan@um.edu.my, suraya_hamid@um.edu.my, ibrahimabaker.targiohashem@taylors.edu.my, ejazahmed@ieee.org

[1]Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya

[2] School of Computing and Information Technology, Taylor's University, 47500 Subang Jaya, Malaysia

[3]Centre for Mobile Cloud Computing Research. Malaysia

Corresponding Authors. E-mail addresses: norjihan@um.edu.my (Norjihan Abdul Ghani), ibrahimabaker.targiohashem@taylors.edu.my (IAT Hashem)

Abstract

Big data analytics has recently emerged as an important research area due to the popularity of the Internet and the advent of the Web 2.0 technologies. Moreover, the proliferation and adoption of social media applications have provided extensive opportunities and challenges for researchers and practitioners. The massive amount of data generated by users using social media platforms is the result of the integration of their background details and daily activities. This enormous volume of generated data known as "big data" has been intensively researched recently. A review of the recent works is presented to obtain a broad perspective of the big social media analytics research topic. We classify the literature based on important aspects. This study also compares possible big data analytics techniques and their quality attributes. Moreover, we provide a discussion on the applications of social big media data analytics by highlighting the state-of-the-art techniques, methods, and the quality attributes of various studies. Open research challenges in big data analytics are described as well.

Keywords: Big data, Social Media, Machine learning, Analytics

1.0 Introduction

The popularity of the Internet and the advent of the Web 2.0 technologies have transformed the contents of the web from publisher- to user-created contents (Alexander, 2006). Such existence has assisted in publishing contents without the needs of programming. Today, interesting topics, reviews, and opinions from Web 2.0 and social media can easily be accessible globally via the Internet in real time. Moreover, the proliferation and adoption of social media have provided extensive opportunities and challenges for researchers and practitioners. More than a billion of people around the world are using social media platforms that generate overwhelming unstructured data in relatively short timescales. The huge amount of data generated by users is the result of the integration of their background details and daily activities in such platform. This massive amount of generated data referred to as "big data" has been intensively researched recently.

The big data from the huge amount of the dataset collected in either structured, semi-structured and/or unstructured format have been researched in various domains, such as health care, astronomy, social web, and geoscience (Hashem et al., 2015). Social media contents, such as tweets, comments, posts, and reviews, have contributed to the creation of big data extensively from either platform providers or different websites (Kwon et al., 2014; Lyu & Kim, 2016). The emergence of big data from social media has brought about a new wave of excitement into the field of artificial intelligence and data analytics. Analyzing social media data using various traditional data mining and machine learning techniques is still an active domain of research. For instance, revealing market research information can be achieved through mining people's opinions that results in improved business decision making (B. Liu, 2015). However, efficient techniques and analytical approaches to handle the ever-growing data generated by various social media applications are of paramount need (Gama et al., 2014). Over the last decade, research related to social media has been increasingly growing, and many algorithms related to artificial intelligence and machine learning have been developed.

This study reviews recent works on big social media analytics using machine learning approaches, provides a broad perspective of the area, and considers challenges and open issues that must be addressed. Perhaps, (Bello-Orgaz et al., 2016; L'Heureux et al., 2017; Qiu et al., 2016) offered the most relevant related work. Other related works include "an introduction to social network data analytics" by (Charu C Aggarwal, 2011), "opinion mining and sentiment analysis" (Pang & Lee, 2008), and networking for big data (Yu et al., 2017).

The rest of the paper is organized as follows: Section 2 introduces the background of social media, big data analytics, and machine learning. Section 3 provides the taxonomy of big data analytics on social media. Moreover, this section discusses the recent advance in machine learning algorithms that have been developed for big data analytics for social media, particularly the application of deep learning, artificial neural networks, fuzzy systems, swarm intelligence, and evolutionary computation. Section 4 provides the discussion on social big media data analytics. Section 5 highlights the research challenges. Section 6 presents the conclusion and future directions.

2.0 Big Social Media Data

Social media (SM) is a set of Internet-based applications that is grounded by the idea of Web 2.0 (E. Gilbert & Karahalios, 2009). SM was initially used around 2004 to describe contents and applications that can be continuously modified and altered by users in many ways through participation and collaboration, rather than traditionally created, prepared, and published by only individuals (Kaplan & Haenlein, 2010). The broad utilization of available software and hardware to access social media platforms over the Internet led to the creation and exchange of user-generated content. Ellison (2007) listed three aspects to define social media that they referred to as "social network site" as web-based services. First, individuals are allowed to create their public or semi-public profile. Second, these individuals are allowed to connect to others to form a network. Last, these individuals are allowed to view and relate to other users and their activities, which are publicized in their network. The terms social media and social media sites have been used interchangeably. In this paper, the term social media refers to any social network sites that have all the three aspects as per Ellison. The examples of social network sites that generate a large amount of unstructured data are Facebook, Twitter, Instagram, LinkedIn, blogs, wikis, and YouTube. Big social media data along with the progress in computational tools have emerged as the key to crucial insights into human behavior and are continually stored and processed by corporations, individuals, and governments (Manovich, 2011).

The most common applications of big data for social media are trend discovery, social media analytics, sentiment analysis, and opinion mining. For instance, social media assists organizations to obtain customers' feedback regarding their products, which can be used to modify decisions and to obtain value out of their business (Katal et al., 2013; Wu et al., 2014). Studies confirmed that most of the existing approaches to big social media data analysis rely on machine learning techniques (Cambria et al., 2013). Some of the most common techniques are classification (Charu C Aggarwal & Zhai, 2012; Reuter & Cimiano, 2012), clustering (Lim et al., 2017; Tang & Liu, 2012), and deep learning (Jansson & Liu, 2017; D. T. Nguyen et al., 2016). Machine learning is a field of artificial intelligence that has been applied in many social media platforms to detect patterns in data. However, dealing with a large amount of data collected from social media with various formats also has brought about some challenges due to the peculiarities of social media such as slang and jargon used in posts. Big data collected from social media

are useless until properly utilized to drive decision making by turning a huge amount of social data into meaningful insights (Gandomi & Haider, 2015). Figure 1 shows the process of big data in social media. First, the data are collected from various social media sources and stored in big data storage technologies that can handle large amounts, such as HDFS, Hbase, and Cassandra. Social media data are noisy and full of irrelevant information for analysis and contain a huge amount of inconsistent data (Wlodarczak et al., 2014). The data integration and cleaning are applied to prepare the big data for processing using technologies such as Spark, Hadoop, and Mesos. Finally, the users can be able to view the result data processing via various end devices, such as computers, servers, and smartphones.

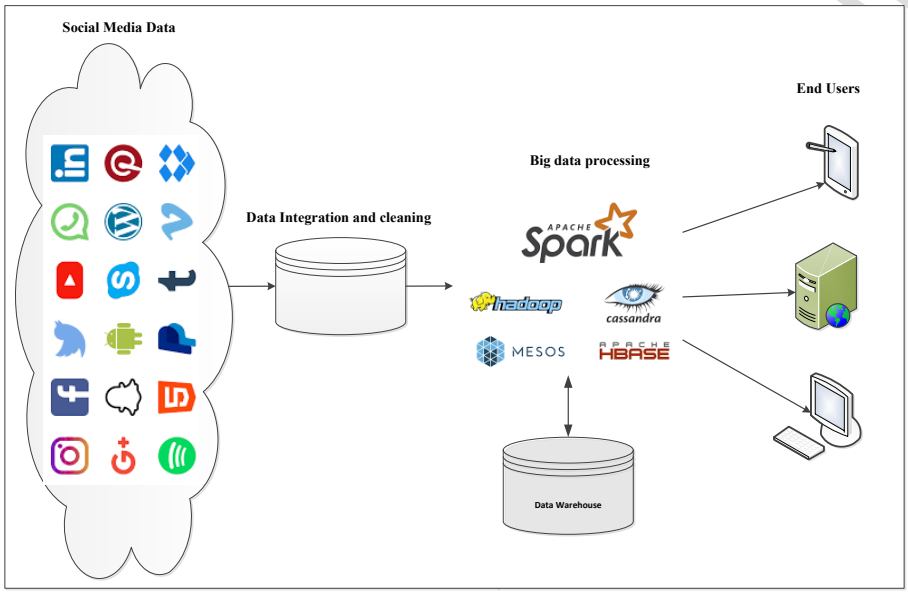


Fig. 1. Big social media data processing

3.0 Big Data Analytics on Social Media Classification

In this section, the taxonomy of big data analytics is discussed on the basis of the previous studies concerning social media data. Many studies focused on big data analysis techniques with social media data (Adedoyin-Olowe et al., 2013; J. Kim & Hastak, 2018; Stieglitz et al., 2018). Big data analytics is capable of handling various dominant research issues using computational intelligences. Thus, big data analytics on social media is categorized into different classes to grasp their characteristics. The taxonomy offered in this review is an effort to address most of the faults and shortcomings of previous works. Figure 2 shows the numerous categories of big data analytics. The classification is based on four aspects: data sources, characteristic, computational intelligence, and techniques. The proposed classification assists in providing a systematic approach to understand the techniques and technologies of big data analytics used in social media data.

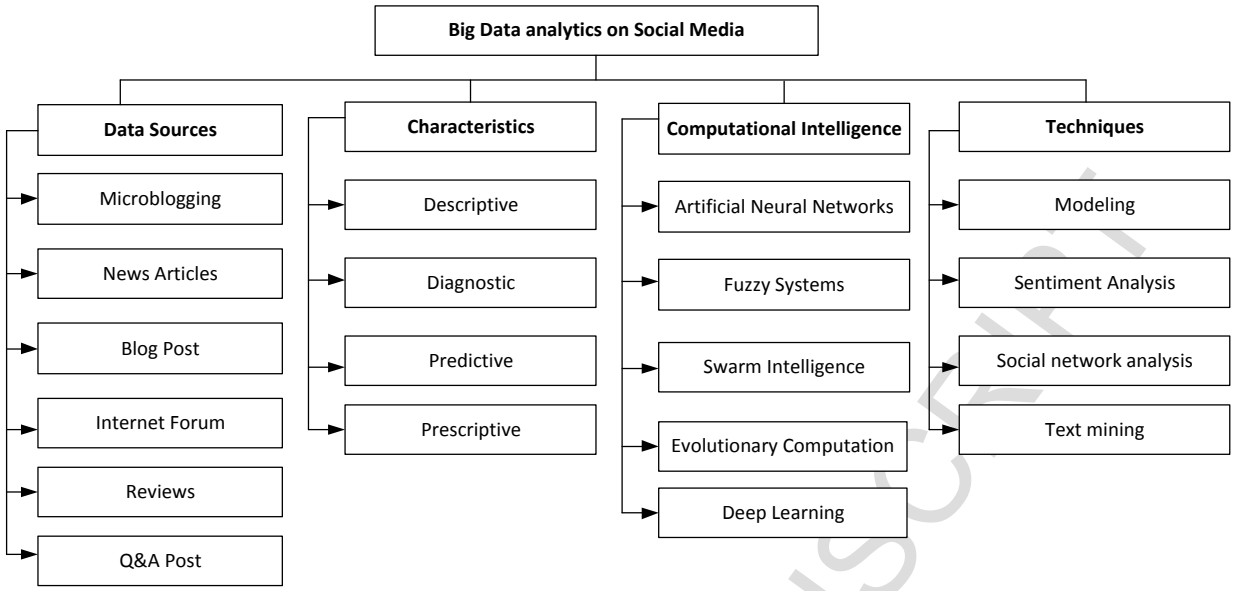


Fig. 2. Classification of big data analytics on social media

Figure 2 illustrates that, for each aspect of the classification, relevant research studies are collected and summarized to provide a background and a deep understanding of the classification. The most related classification found in the literature is discussed in (Ahmed et al., 2017), which focused on the big data analytics in the Internet of things. However, this study concerns with the use of machine learning approaches for big data analytics in social media.

3.1 Social media data sources

The widespread popularity of social network has dramatically increased the rate of various data creation. Such increased data generation has drawn attention to the needs of real-time analytics. The variety of data is the different types of structural heterogeneity in a dataset. These different types of datasets are structured, semi-structured, and unstructured data, such as Microblogging, news articles, blog post, internet forum, reviews, and Q&A posts. For instance, M. Kim and Leskovec (2012) used AddHealth, Egonet, Facebook100, and WebKB datasets, which contain networks and node features to evaluate their model. Moreover, Spagnuolo et al. (2014) used Bitcoin OTC trust-weighted signed network datasets, which allows the monitoring of the Bitcoin economy through tracing users' money using attached identities. The massive amount of data generated by users is the result of the integration of their background details and daily activities in this platform. All the statuses, tweets, comments, posts, and reviews are the user-generated content. User-generated content is a type of data that typically refers to images, text, and videos. This content comes from regular people and not necessarily in a standard form. Therefore, various quality distributions of user-generated contents occur, which range from high-quality to low-quality things as the data generated by social media sites are naturally fuzzy and unstructured. All these data may incorporate the users' personal opinion, behaviors, and thoughts, which makes the task of extracting high-quality information from such data becoming increasingly important. It is a plentiful area to be discovered for businesses and researchers due to the availability of user-generated content that possibly encapsulates useful and high-quality information.

Selecting relevant literature reviews is an important task. The selection process is ensured to be comprehensive, unbiased, and an extensive systematic search for all literature pieces that meet the preferred criteria of this paper. The initial search is performed to retrieve literature that is published in credible journals to ensure quality. The selected journals are from a high-impact journal that is referred from Journal Citation Report (JCR), which is annually produced by the Science and Scholarly Research division of Thomson Reuters.

The search process is executed by using the keyword parameter "*social media data analysis*" through *Publish or Perish 4* software by Harzing (2007) for every selected journal. This software allows an efficient search of the literature by analyzing research publications and their citations. The retrieval is initiated in the "General Citation Search" category by entering all the specified "keyword parameter" (as mentioned above) in "All of The Words" field while at the same time restricting the selection of publication year between 2011 and 2017. The "data source" field is set to be Google Scholar. Google Scholar is chosen to provide access to a broad range of scholarly literature.

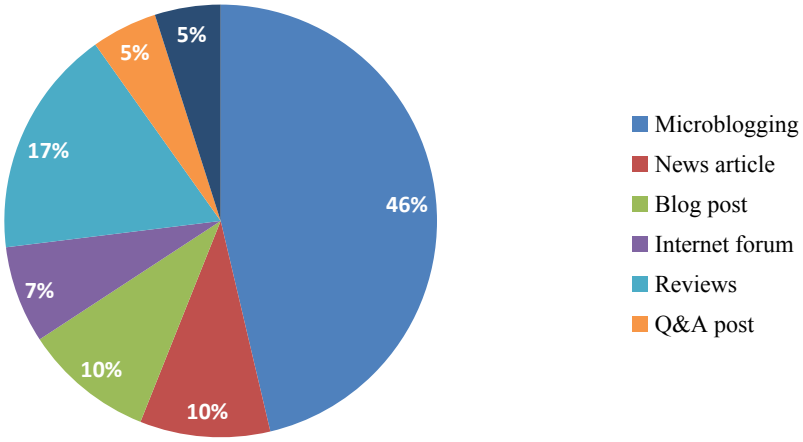


Fig. 3. Social media data sources

Figure 3 depicts that 46% of the current studies perform analysis on *tweets* produced in microblogging. A total of 17% of these studies perform analysis on *reviews* by users, such product and movie reviews. The comments are also considered the opinions and views tracked in textual content. Moreover, 10% of these studies use *blogpost* as their data sources, while *news articles* and *forum posts* are used for analysis by 9% and 7% of the selected studies, respectively. A total of 5% of the studies use Q&A posts (from question and answers sites) as their data source. Similarly, tags in sites, such as Tumblr and Flickr, are also used by 5% of the selected studies. Table 1 shows the summary of social data sources that are used during the analysis in different domains. Microblogs have become a major source of information for users (Hong et al., 2011). Thus, most of the research focuses on microblogs as the main data sources for data analysis, especially in the social domain, such as sentiment, classification, and text analysis.

Table 1. Summary of social data sources

Data Sources	Type of Data Analysis	Discipline	Reference

Microblogs	Sentiment analysis	Computer Science (Neurocomputing)	(Li et al., 2016)
	Event detection and analysis	Medical	(W. Liu, Luo, Gong, et al., 2016)
	Principal components analysis	sociolinguistic and natural language processing	(Yuan Huang et al., 2016)
	Regionalization with Constrained Clustering and Partitioning (REDCAP)		
	Text analysis	Social	(Jaidka et al., 2016)
	Classification analysis	Social	(A. X. Zhang & Counts, 2016)
	Text analysis	Social	(Fast et al., 2016)
	Scaling and Opinion mining	Retailing	(Tse et al., 2016)
	Sentiment analysis	sociolinguistic and natural language processing	(Aisopos et al., 2016)
	Opinion mining, text categorization	Sport	(Barnaghi et al., 2016)
	Sentiment analysis	Sport	(Hoeber et al., 2016)
	Opinion mining	Sport	(Samariya et al., 2016)
	Spatiotemporal analysis	Social	(Dong et al., 2015)
	Spatiotemporal analysis	Social	(Cao et al., 2015)
	Topic model analysis	Social	(Ito et al., 2015)
	Temporal analysis	Social	(Azarbonyad et al., 2015)
	Text analysis	Social	(Lovinger & Valova, 2015)
	Predictive Analysis	Social	(Yeon & Jang, 2015)
	Temporal analysis	Social	(W. Liu, Luo, Gong, et al., 2016)
Facebook	Difference in Differences (DID) analysis	Social	(Grinberg et al., 2016)
	Sentiment analysis	Law	(Nouh & Nurse, 2015)
	Temporal analysis		
	Sentiment analysis	Social	(Gao et al., 2015)
Wikipedia	Temporal statistical analysis	Social	(Do et al., 2016)
Weibo	Filtering analysis	Social	(Xu et al., 2017)
	Multi-modal mining (semantic, spatial temporal and visual)		
	Sentiment analysis	Aviation	(S. Chen et al., 2016)
Flickr	Classification analysis	Social	(Qian et al., 2015)
Google+	Performance analysis	Computer science	(Santos et al., 2016)

3.2 Characteristic

The characteristic of analytics in big data can be divided into four types, namely, descriptive, diagnostic, predictive, and prescriptive. Table 2 shows tht summary of the characteristic analytics on social media.

3.2.1 Descriptive analytics

In data processing, descriptive analytics is the initial phase that provides all the necessary historical/past data needed to provide valuable information. Moreover, it can be used in facilitating further data analysis (Tan et al., 2014). By a detailed understanding of the successful and failure parts of past data, historical data are provided. Descriptive analytics provides techniques on analyzing data based on this information. Also called as "post-mortem analysis," descriptive analysis is mostly used in organizations for reporting of events, such as monitoring of sales, departmental, and finance. The descriptive models can quantify, identify, and categorize various relationships/connections in data (Simpao et al., 2014). In addition, the descriptive modeling tools can be used to improve additional models that can simulate a considerable number of customized agents and make predictions.

3.2.2 Diagnostic analytics

The diagnostic analytics is an improved type of analytics characterized by techniques, such as data discovery, drill-down, data mining, and data correlations. Diagnostic analytics scrutinizes data or order to answer questions, such as "For what reason did it happen?". This analytics thoroughly investigates data to know the detailed behavior and causes of events (Wang et al., 2016). The diagnostic analytics provides a chance of fast data comprehension and quick attempts to some critical workforce questions. Cornerstone View is an example of diagnostic analytics that provides the speediest and easiest path for management/organizations to acquire valuable knowledge with respect to complex issues. It uses data visualization as an interactive tool that enables managers to enable effortlessly search and filter individuals by integrating data (Wamba et al., 2015). For instance, managers can uncover significant knowledge about their employees when applying for critical position and the succession metrics and the performance of the exisiting employees.

3.2.3 Predictive

The predictive analytics transforms data into profitable and actionable information (Ayhan et al., 2013). This analysis utilizes data to decide the possible future result of an event or a possibility for a certain situation to occur. Predictive analytics involves different statistical methods that range from modeling, machine learning, and game theory for analyzing present and past facts to forecast future events. The historical data found in predictive models are used in businesses to identify risks and opportunities. The model identifies the relationships between various factors to permit risk assessment based on a specific set of condition, which will guide the decision-making process. Predictive analytics has three major stages (Kowalczyk & Buxmann, 2015), namely, modeling (predictive), decision examination and optimization, and transaction profiling. The optimization of customers' relationship management system is an example of predictive analytics. It allows the organization to analyze all their customers' data and to predict their customers' behavior. In sum, predictive analytics leads to high benefits and several strong customer relationships (Delen & Demirkan, 2013).

3.2.4 Prescriptive analytics

Prescriptive analytics can similarly propose decision opportunities on exploiting a future opportunity or lessen future risks. This analytics demonstrates the effects of every decision opportunity. Practically, prescriptive analytics can constantly and spontaneously process new data to enhance the precision and to offer good decision options (Song et al., 2013). A prescriptive approach is an intensive approach that examines possible decisions, the connections among decisions, and the impacts and results of these decisions, which finally propose an ideal option in real time. The efficiency of predictive analytics depends on the sufficiency of the decision model in capturing the effect of the decisions being analyzed (Song et al., 2013). Furthermore, optimization, game theory, simulation, and decision techniques are the specific approaches employed in prescriptive analytics.

Table 2. Summary of the characteristic of big data analytics on social media

Features	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Scope	What has already happened?	For what reason did it happen?	What will happen in next future? What trends will continue?	How to achieve the best outcome for any given condition?
Tools	Dashboards, Charts and Key Performance Indicators (KPIs)	Interactive data visualization	Statistical Methods, Data mining, and Modeling	Statistical optimization techniques, game theory, and Simulations
General examples	Monitoring of sales, Departmental operations, and Finance	Cornerstone View	Decision examination, Optimization, and Transaction profiling	Propose decision opportunities
Application Areas	Netflix uses data mining to determine the correlations between various events	Health-care centers use interactive data visualization for speedy and easy path for management to acquire important knowledge about their employees and solve complex workforce issues	ING uses it to analyze all their customer's data and enable them to also predicts their customer's behavior	Amazon.com utilizes price optimization based on demand to increase the online shopping revenues.

3.3 Computational intelligence

Social media platforms have gained popularity of behavior-rich resources and interactive platforms, which offers challenges and opportunities in big data analytics (Banerjee & Agarwal, 2012). Moreover, these platforms require an active use of computational intelligence to help understand opinions (Banerjee & Agarwal, 2012). Computational intelligence is a new research field, which is defined by (Fulcher, 2008) as the “adaptive mechanisms that enable or facilitate intelligent behaviors in complex and changing environments”. Figure 3 shows various classifications of the computational intelligence approaches that

are employed on processing a large amount of social data. The classification, as mentioned by (Fulcher, 2008) in his book "Computational Intelligence: An Introduction," is based on artificial neural networks (ANN), fuzzy systems, swarm intelligence, evolutionary computation, and deep learning.

3.3.1 Artificial neural networks

Artificial neural networks is inspired by the brain modeling called neuron or perceptron, which is effluenced by its internal state activation received from input (Fulcher, 2008; Schalkoff, 1997). (Straton et al., 2017) attempted to improve the information dissemination on social media for health-care orgnizations by reducing clutter and noise. The study used Facebook to predict the popularity of heath-care posts based on the eleven characteristics of the post. (Lohokare et al., 2017) presented a mechanism for financial transaction and social media data collection for credit score calculation. The authors used the artificial neural network technique in obtaining the insight into general social status to claculate the credit scores. (Ghiassi et al., 2016) applied the artificial neural network technique in dynamic architecture and supervised feature engineering on brands using Twitter sentiment analysis based on a targeted approach. The solution is to overcome the problem with the brand-related tweet sentiment class distribution and the unique characteristics of the Twitter language. The authors managed to illustrate the usefulness of the proposed approach on Twitter data sets.

3.3.2 Fuzzy systems

(Shanthi & Pappa, 2017) used the gas/liquid two-phase flow with principal component analysis to improve the classification accuracy of its flow pattern using support vector machine and fuzzy logic algorithms. Six types of flow patterns for the video have been utilized to covert 2D images for analysis. An image processing technique is used to extract textural and shape features as the inputs to various classification schemes (Shanthi & Pappa, 2017), such as fuzzy logic. The results corroborate that the performance of the support vector machine and fuzzy logic algorithms with fewer features offer better accuracy compared with other machine learning algorithms. These are also computationally less intensive than the other two current techniques (Shanthi & Pappa, 2017).

3.3.3 Swarm intelligence

Swarm intelligence is an effective problem-solving technique in artificial intelligence (Kennedy, 2006). This technique employs the concept of swarm, such as ants, bees, and wasps, and intelligence based on the structured collection of interacting organisms or agents (Blum & Li, 2008; Engelbrecht, 2006; Fulcher, 2008). Recently, this technique has been used in analyzing humans' social behavior pattern. For instance, Lv et al. (2016) used a social media platform to provide study on the complicated behavior of human social by proposing a Twitter optimization algorithm based on swarm intelligence. The proposed algorithm offers a good solution with respect to most of the real parameter optimization problems. Moreover, (Hu et al., 2017) introduced a new indexing method named hybrid quantum swarm intelligence for social networks. The proposed algorithm is based on the fluctuation detectionand optimal weight algorithms for link prediction. Moreover, nature-inspired theory based on swarm intelligence is proposed by (Banerjee & Agarwal, 2012) to model the behavior of the group of users in various blogs. The idea is to be able to predict the event behavior of a large popoulation during a training phase in an acurate mannar in the future after observing their interactions.

3.3.4 Evolutionary computation

Evolutionary computation (EC) is a subfield of the artificial intelligence that is based on natural evolution and has been successfully used to solve various optimization problems. The concept of ECs has been drawn from the idea of survival of the fittest, natural selection, reproduction, mutation, competition, and symbiosis (Fulcher, 2008). Genetic algorithms are among the most useful approaches for multi-objective nonlinear discrete optimization problems. Various implementations of genetic algorithms have been used over the years, such as NSGA-II (non-dominated sorting genetic algorithm) (Deb et al., 2002) and SPEA2 (strength Pareto evolutionary algorithm) (Zitzler et al., 2001). For instance, distributed genetic algorithms have been proposed by (Hajeer et al., 2014) for the detection of communities in networks. The authors introduced an alternative way to encode the network chromosomes, which greatly optimizes the memory use and computations that offer an efficient and scalable framework. Takagi (2001)

3.3.5 Deep learning

The state-of-the-art literature on the applications of computational intelligence algorithms in big social media datasets shows that the most common algorithms are the deep neural networks (DNN) (Havaei et al., 2017) and large-scale recurrent ANN. However, these classes of ANN are susceptible to limitations as follows: the DNN is mostly trained using first-order stochastic gradient descent, e.g., the backpropagation training algorithm, which is difficult to parallelize due to its serial execution; the high number of parameters involved in the training of DNN on big data; slow convergence speed, especially for big data; high computational cost; vanishing of gradients; explosion; and the problem of inter-processor communication cost and bottlenecks incurred by the parallelisation algorithms. Previous studies using DNN for big data analytics affirm that the researchers mainly rely on the backpropagation algorithm for the training of the DNN for big data analytics; however, the backpropagation algorithm is susceptible to falls in local minima. This also has a slow velocity for convergence to the optimum solution during learning. Currently, the long convergence time of the DNN limits the use of DNN to high-cost servers or platforms with multiple GPU cores due to the complex structure of the DNN. Many other studies focused on computational intelligent with respect to social media. For instance, (Fan et al., 2017) applied DNN on feature extraction. The solution is based on a low-level structure for image processing to improve the preformance of learning using input data. The analysis of the features is used to improve the edge-preserving filters of the image processing. The result confirms that edge-preserving filters can solve problems, such as noise amplification, edge blurring, and halos.

3.4 Techniques

The big data analytics techniques in social media context are related to natural language processing, sentiment analysis, social network analysis, and news analytics. This section presents important techniques for big data analytics with respect to social media data analysis. These techniques play an important role in improving the business and decision-making through analysis (Shanthi & Pappa, 2017).

3.4.1 Modeling

Social media analytics deals with managing and evaluating informatics tools for social media data collection, monitoring and analyzing (Elkaseh et al., 2016). It is extracting process that provides a suitable pattern for data analyses during conversations and interactions. The analytical processes in social media analytics is divided into three (3) phases, namely, capturing phase, understanding phase and presentation phase. Social media analytics involves the use of different modeling and analytical techniques from

various fields (T. H. Nguyen & Shirai, 2015). In addition, social network graph are used in social network analysis to provide a detail understanding of its fundamental structure, connections, and theoretical properties (C. H. E. Gilbert, 2014). Its classified the comparative importance of various nodes inside the network.

3.4.2 Sentiment analysis

The sentiment analysis is a class of NLP, text analysis, and statistics. The idea is to find the sentiment of the text by classifying them into positive, negative, or neutral. This analysis is usually used for binary decision making, i.e., users like or dislike something, or the product is good or bad (Ohbe et al., 2017). Sentiment analysis, which is also named as opinion mining, includes categorizing consumer attitudes, emotions, and opinions of a company's product, brand, or service. In social media, sentiment analysis has various uses. For instance, this analysis can be applied to identify the feelings of consumers in a marketing and customer service department, which results into uncovering whether consumers are satisfied or dissatisfied with a product (Povoda et al., 2017).

3.4.3 Social network analysis (SNA)

Social network analysis is based on finding the social relationships between different users from social media using network theory of nodes and connections (Serrat, 2017). Social network analysis has emerged as a key technique in social media applications, such as fraud detection and sociology. This analysis has also gained a significant following in medicine, anthropology, biology, and information science. SNA has become a popular topic of speculation and study. SNA has undergone a renaissance with the ubiquity and quantity of content from social media, web pages, and sensors (Cybenko, 2017). Social networks are embedded in many sources of data and at many different scales. Social networks can arise from information in sources, such as text, databases (Campbell et al., 2013), sensor networks, communication systems, and social media (Evans, 2017).

3.4.4 Text mining

Text mining emerges as the most popular techniques in social media for information extraction from numerous types of unstractured contents, such as text, images, and multimedia (Charu C. Aggarwal & Wang, 2011). A detailed survey on text mining on social media is presented by (Reddick et al., 2017; Salloum et al., 2017). In addition, (McCaig et al., 2018) used text mining techniques to compare disorder communities data related to their eating habits based on fitness tracking technology. Three types of eating related to disorder data are collcted within three years from comments posted in social media. The data are composed of six subreddits, fitness, and weight management. The results validate that subreddits with pre-disorder are less recovered compared with eating disorder subreddits with the highest frequency of fitness tracker terms. (Cheng et al., 2017) explored whether computerized language analysis methods can be utilized to assess one's suicide risk and emotional distress in Chinese social media. The framework for social media content is proposed by (Thomaz et al., 2017). The authors used emprical study to test the framework during the FIFA World Cup 2014.

4.0 Discussion

The literature on the applications of social big media data analytics, as shown in Table 3, highlighted the state-of-the-art techniques, methods, and the quality attributes that offer an efficient performance on analyzing social media data. The selected publications listed in Table 3 are based on the contribution effort on the domain of social media analysis using machine learning, metaheuristic, and evolutionary

approaches. Moreover, various categories of the specified purpose of conducting analysis for social media data have been identified on the basis of the reviewed papers. The main focuses of most studies are users' emotion classification, information detection, spatiotemporal, clustering, and performance analysis. Thus, social media plays a vital role on modern data analysis for achieving the meaningful information. Sentiment analysis can be defined as a process of determining favorable or unfavorable opinions toward a specific subject in a texts (Nasukawa & Yi, 2003). On the basis of the review process, sentiment analysis classificaion approaches have three levels, which are aspect, document, and sentence. The document level is when the prediction of sentiment is made to the entire document based on analysis done to the parts of the document. The sentence level is when the analysis of sentiment is done to the sentence parts to extract the sentiment for that particular sentence. Document and sentence are considered the basic unit of information for each classification. (Bing, 2015) affirmed that no fundamental difference exists between these two as a sentence can be regarded as a short document. Differently, in the aspect level, the prediction of sentiment polarity is based on the aspect of a particular subject rather than on the literal meaning of a word to be analyzed. Furthermore, some features have been identified to correspond to the information detection process, which are mentioned in Table 3. They are spatiotemporal—the information about location and occurrence time, temporal—the information about occurrence time, social link—the information about users' network and behavior, semantic—the information related to the meaning of words or terms, sentiment—the information related to users' emotions, and term co-occurrence—the information related to frequency or words or terms. Information extraction from text can be defined as the process of drawing out useful or intended data from text. With regard to the discussion in Hogenboom et al. (2011), information extraction in text is defined as the discovery of a set of empirical observations from text that is grouped into three main approaches: (i) data driven, (ii) knowledge driven, and (iii) hybrid. A data-driven approach does not consider the semantic aspect of the information. This approach requires many data but little domain expertise. Differently, a knowledge-driven approach considers a semantic aspect of text by using linguistic, lexicographic, and human knowledge for the extraction process. This aspect requires considerable domain and expertise knowledge but only little data. A hybrid approach is the combination of both aspects.

There are several aspects of research in the text analysis of social media data that have been covered, such as social media data categories, aims, approaches and methods involved, and data platform used for analysis. Social media data analysis is an up-and-coming area to be researched.

Table 3. A comparison of big data analytics on social media based on their approaches, techniques, and quality attribute.

Objective	Contributions	Approach	Data Analysis	Technique/methods	Quality Attribute	Reference
User emotion classification over short text	multi-label maximum entropy (MME) model	Principle of maximum entropy (ME), Co-training algorithm Generalized MME model (gMME)	Sentiment analysis, social emotion classification, and short text modeling	Experimental	Influence of the iteration number	(Li et al., 2016)
Discovering the core semantics of event	Markov random field based method	semantics collaborative computation for learning association relation distribution	Event detection and analysis	Experimental	Accuracy	(W. Liu, Luo, Gong, et al., 2016)
Generate multi-modal summary	Multi-Modal Storytelling of Urban Emergency Events	Crowdsensing based multi-modal method	Filtering analysis Multi-modal mining (semantic, spatial temporal and visual)	Case studies	Accuracy	(Xu et al., 2017)
Sentence ordering	Cognitive memory-inspired sentence ordering	Semantic Markov Random Field	Comparative analysis	Experimental	Accuracy	(W. Liu, Luo, Xuan, et al., 2016)
Investigate the usefulness of a multimodular approach	Multimodular systemfor text normalization.	text normalization approach	Performance analysis	Experimental	performance	(Schulz et al., 2016)
To provide awareness through emergency and disaster events	A transparency-based spatial context preserving technique	Visual analytics	Multiple scales analysis.	Experimental	accuracy	(J. Zhang et al., 2016)
Classifying groups based on standard community quality measures.	Community evaluation measure.	COMODO algorithm	Descriptive mining	Experimental	accuracy	(Atzmueller et al., 2016)

To merge affective information extracted from multiple modalities.	Multimodal sentiment system.	Feature- and decision-level fusion methods	Sentiment analysis	Experimental	Accuracy	(Poria et al., 2016)
To examine the query side	Provide a better solution for social media data analytics.	A query expansion approach	Temporal statistical analysis	Experimental	Accuracy	(Do et al., 2016)
Using Twitter to Identifying fine-grained stories within a wider trending topic.	Text-based similarity calculation metrics	Text Classification and clustering	Text analysis	Experimental	Accuracy	(Jaidka et al., 2016)
To better understand the changing needs and preferences of contributors	Level of Engagement Before and after Posting	observational data analysis	Difference in Differences (DID) analysis	Observation	Changes	(Grinberg et al., 2016)
To study rapid spread of anti-abortion policy change	characterize people's expressions of opinion on abortion and show how these expressions align with the policy change	Correlation approach	Classification analysis	Experimental	Accuracy	(A. X. Zhang & Counts, 2016)
To create new lexical categories on demand.	Tools for generating and validate new lexical categories	Modern NLP techniques with the transparency of dictionaries (LIWC)	Text analysis	Experimental	Accuracy	(Fast et al., 2016)
Detection and classification of urban events	method for classifying urban events.	Profile mining and events classification	Classification analysis	Experimental	Accuracy	(Sato et al., 2016)
Explore data locality to reduce processing costs	Performance improvements of high-level programming interface	Comparative approach	Performance analysis	Experimental	Performance	(Santos et al., 2016)
Explore the attitudes of UK consumers by identifying the hidden information in tweets	A framework which can assist industry practitioners in managing social media data.	A comprehensive data analysis framework	Scaling and sentiment analysis	Experimental	Accuracy	(Tse et al., 2016)

To examine extensive experiment results with a multilingual corpus of manually annotated posts	improving the classification accuracy compared to the current State of the Art.	the supervised machine learning model	Sentiment analysis	Experimental	Accuracy	(Aisopos et al., 2016)
Examine a positive or negative sentiment on Twitter posts	a correlation between twitter sentiment and events that have occurred	Correlation approach	Sentiment analysis, text categorization	Experimental	Accuracy	(Barnaghi et al., 2016)
To understand better the passengers and improve customer relationship management	a comprehensive profile for travelers.	Modeling passenger's value	Sentiment analysis	Experimental	Accuracy	(S. Chen et al., 2016)
To conceptualize, model, analyze, explain, and predict social media interactions	A new approach to big data analytics called social set analysis	Comparative approach	fuzzy set-theoretical sentiment analysis crisp set-theoretical interaction analysis	Experimental	Accuracy	(Mukkamala et al., 2014b)
To support the interactive discovery of emergent themes	Vista supports an exploratory analysis and dynamic filtering of a large collection of tweets,	Visual analytics	Sentiment analysis	Experimental	Accuracy	(Hoeber et al., 2016)
To test emotions of Indian Cricket Lovers.	Captures emotions for big data analytics	Lexicon and machine learning method	Sentiment analysis	Experimental	Accuracy	(Samariya et al., 2016)
Event detection	A novel algorithm to compute a data similarity graph.	clustering approach	Spatiotemporal analysis	Experimental	Accuracy	(Dong et al., 2015)
To harness massive, un-structured location-based social media (LBSM) data	a scalable computational framework	Space-time trajectories (or paths)	Spatiotemporal analysis	Experimental	Accuracy	(Cao et al., 2015)
to effectively extract information from multiple modalities.	Multimodal information extraction agent	Ensemble feature extraction approach	Multimodal sentiment analysis	Experimental	Accuracy	(Poria et al., 2015)

cross-domain data analysis	algorithm for cross-domain data analysis.	nonparametric Bayesian dictionary learning model	Classification analysis	Experimental	Accuracy	(Qian et al., 2015)
Examining the value of organizational documentation in evaluating social media activism	The Value of Organizational Texts in Network Analysis	System of Texts	Rhetorical analysis	Experimental	Accuracy	(Trice, 2015)
to predict and mitigate customer churn.	Extracting dominant features from user complaints and Web data for churn prediction	a maximum entropy-based approach	Sentiment analysis (temporal analyses)	Experimental	Accuracy	(Das et al., 2015)
To automatically assess the credibility of information	Effective features for classifier on the basis of data analysis results	Latent Dirichlet Allocation (LDA) model	Topic model analysis	Experimental	Accuracy	(Ito et al., 2015)
Time-Aware Authorship Attribution	approach to estimate the dynamicity of authors' word usage	time-aware language models	Temporal analysis	Experimental	Accuracy	(Azarbonyad et al., 2015)
Explore the gathering of text data from online social media	The algorithm to generate rules for predictive text	association rules based on Apriori algorithm	Text analysis	Experimental	Accuracy	(Lovinger & Valova, 2015)
Identifying Key-Players in Online Activist Groups	Approach to identify the interaction among users activist in Facebook.	Social Network Analysis (SNA) techniques	Sentiment analysis Temporal analysis	Experimental	Accuracy	(Nouh & Nurse, 2015)
To retrieve similar event patterns	Predictive event patterns	Topic composition for text data	Predictive Analysis	Experimental	Accuracy	(Yeon & Jang, 2015)
to analyze spatiotemporal patterns	a scalable and distributed geographic information system (DART)	s reasonable pre-splitting and hash techniques	Temporal analysis	Experimental	Accuracy	(H. Zhang et al., 2015)
Handling, storing and analyzing big data	Prevention and decision support for better health.	Apache Hadoop framework and Mahout	Sentiment analysis	Experimental	Accuracy	(Cunha et al., 2015)

differences across privacy levels and demographic factors	Find relationships among various attributes using distinct sentiments.	Identify patterns of text and metadata using mining technqueis.	Sentiment analysis	Experimental	Accuracy	(Gao et al., 2015)
Story detection	Novel algorithm to extract information from triplets	generalized concepts and relations (Clustering)	Text analysis	Experimental	Accuracy	(Ceran et al., 2015)
Forecasting model	Demand Forecasting Models for Medicines	VARX model	Topic Trend Analysis	Experimental	Accuracy	(W. Kim et al., 2015)
Multiple classifiers for social event classification	BMM-SLDA algorithm	boosting weighted sampling strategy	Classification analysis	Experimental	Accuracy	(Qian et al., 2014)
To extract useful interaction behavior of Twitter users.	Extended fuzzy association rules algorithm (MASS-FARM),	Fuzzy association rule mining	Associations analysis	Experimental	Accuracy	(Fu & Shen, 2014)
To provide a management of the humans and robots interaction using microblogging	a novel approach to social data analysis	Natural language processing	Semantic analysis	Experimental	Accuracy	(Bell et al., 2014)
To study how different sentiment scopes, complement each other.	Meta-level sentiment models	Data mining	Sentiment analysis	Experimental	Accuracy	(Bravo-Marquez et al., 2014)
To cluster a Weibo Message	A new approach to cluster Weibo data	Data mining	Clustering analysis	Experimental	Accuracy	(G. Zhang et al., 2014)
Understanding Students' Learning Experiences	a workflow.	Data mining	Qualitative analysis Classification analysis	Experimental	Accuracy	(X. Chen et al., 2014)
observing and analyzing human behaviors during the Hurricane Sandy disaster	a scalable social media data analytics system	spectral clustering algorithm	Clustering analysis	Experimental	Accuracy	(Yin Huang et al., 2014)

to study the characteristics of retweets	a method to model retweeting	Simulation approach	Frequency distribution analysis	Experimental	Frequency	(Lu et al., 2014)
Computational approaches to social media analytics	unified modelling approaches to social data	a formal model based on fuzzy set theory	Sentiment Analysis	Experimental	Accuracy	(Mukkamala et al., 2014a)
to identify and characterize students' comprehension problems around knowledge artifacts	general methods of artifact analysis have to be combined with adequate indicators (such as signal concepts) and representations (such as combination diagrams)	Feature extraction	Semantic text analysis	Experimental	Accuracy	(Erkens et al., 2014)
Cross-lingual DC for short-text documents.	Cross-lingual Short-Text Document Classification	Classification approach	Text analysis	Experimental	Accuracy	(Faqeeh et al., 2014)
Followee recommendation	algorithm for recommending followees in Twitter	Content-based and exploration of the topology of the network	Text analysis	Experimental	Accuracy	(Armentano et al., 2013)
Profiling Twitter users	a hybrid text-based and community-based method for the demographic estimation of Twitter user	Web mining User profiling	Text analysis	Experimental	Accuracy	(Ikeda et al., 2013)
Irony detection Negation	a new model of irony detection	Figurative language processing	Text analysis	Experimental	Accuracy	(Reyes et al., 2013)

5.0 Research Challenges

Big data analytics on social media bring significant advantages to companies and individuals regarding good decision making. However, several challenges remain unaddressed. Solving some problems related to big data requires concentrated efforts from the big data research communities, companies, and government to overcome such challenges.

1) ***Data Quality:*** User-generated content is a type of data that is typically in the form of text, images, and videos. This content comes from regular people and is not necessarily in standard form. Hence, the various quality distributions of user-generated contents range from very high-quality to low-quality things as the data generated by social media sites are naturally fuzzy and unstructured (Kwon et al., 2014). All these data may incorporate users' personal opinion, their behaviors, and thoughts, which makes the task of extracting high-quality information from such data increasingly important. It is a plentiful area to be discovered for businesses and researchers due to the availability of a user-generated content that possibly encapsulates useful and high-quality information.

2) ***Data Locality****:* Data locality is the major challenge associated with different types of data (Leskovec, Rajaraman, & Ullman, 2014). In addition, all machine learning algorithms assume that datasets are located in the memory (Parker, 2012). However, this assumption is not applicable for big data because the complete data cannot fit into a memory. The data are usually distributed across different files residing in various physical locations. Big data is entirely different from the conventional machine learning that only transfers data to a computing location. Thus, for a successful big data system, the issue data locality of poor processing latency and network traffic must be adressed.
3) ***Velocity****:* Most machine learning techniques rely on data availability, which implies that learning cannot start when the entire dataset is unavailable (Gu et al., 2015). However, achieving data availability will not be possible in the case of data streaming because of the continuous arrival of new data. Big data velocity deals with the speed and the rate at which the data are procssed. The rapid growth in technologies, such as mobile devices and real-time sensor nodes, leads to rapid communication between our environment and other smart devices. Therefore, big data velocity during the development process is must be considered.
4) ***Data Availability****:* Machine learning techniques solely rely on data availability. Therefore, in machine learning, the entire dataset must be available before learning can begin. Moreover, a model normally learns from the training set and later performs the learned tasks, such as prediction and classification. However, this scenario cannot be achieved in the case of data streaming where new data are continually arriving (Hu & Zheng, 2015).
5) ***Natural language processing(NLP):*** NLP is an intelligent process of analysis and understands the meaning of the texts from the human language. Many tasks can be performed through the development of structure knowledge, such as relationship extraction, speech recognition, automatic summarization, topic segmentation, sentiment analysis, and entity recognition (Ismail et al., 2016). Social media analysis is a good example of NLP that allows computers to process the unstructured texts collected from social media applications. This processing can be observed

in forms, such as topic extraction, relationship extraction, named entity recognition, automatic text summarization, and stemming. NLP is commonly used for text mining, machine translation, and automated question answering (Manning & Schütze, 1999). NLP is characterized as a challenging problem in the area of big social data (Syeda et al., 2017). The human language is rarely precise or plainly spoken. In comprehending the human language, not only the words but also the context and their relation to each other must be understood. Despite language being one of the easiest things for humans to learn, the ambiguity of the language is what makes natural language processing a difficult problem for computers to master (Manning & Schütze, 1999).

6.0 Conclusion

Big social media data along with the progress in computational tools have emerged as the key to crucial insights into human behavior and are continually stored and processed by corporations, individuals, and governments. Frist, we investigate the recent work on big social media analytics and provide a broad perspective of the area. Second, we classify the literature based on important aspects, such as data sources, characteristics, computational intelligence, and techniques. Third, we compare possible big data analytics techniques and their quality attributes. Fourth, we identify the opportunities with respect to big data analytics on social media. Finaly, we describe the open research challenges in big data analytics.

References

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

Aggarwal, C. C. (2011). An introduction to social network data analytics. *Social network data analytics*, 1-15.

Aggarwal, C. C., & Wang, H. (2011). Text Mining in Social Networks. In C. C. Aggarwal (Ed.), *Social Network Data Analytics* (pp. 353-378). Boston, MA: Springer US.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*: Springer Science & Business Media.

Ahmed, E., Yaqoob, I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., & Vasilakos, A. V. (2017). The role of big data analytics in Internet of Things. *Computer Networks*.

Aisopos, F., Tzannetos, D., Violos, J., & Varvarigou, T. (2016). *Using n-gram graphs for sentiment analysis: an extended study on Twitter.* Paper presented at the Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on.

Alexander, B. (2006). Web 2.0: A new wave of innovation for teaching and learning? *Educause review, 41*(2), 32.

Armentano, M. G., Godoy, D., & Amandi, A. A. (2013). Followee recommendation based on text analysis of micro-blogging activity. *Information systems, 38*(8), 1116-1127.

Atzmueller, M., Doerfel, S., & Mitzlaff, F. (2016). Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences, 329*, 965-984.

Ayhan, S., Pesce, J., Comitz, P., Sweet, D., Bliesner, S., & Gerberick, G. (2013). *Predictive analytics with aviation big data.* Paper presented at the Integrated Communications, Navigation and Surveillance Conference (ICNS), 2013.

Azarbonyad, H., Dehghani, M., Marx, M., & Kamps, J. (2015). *Time-aware authorship attribution for short text streams.* Paper presented at the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Banerjee, S., & Agarwal, N. (2012). Analyzing collective behavior from blogs using swarm intelligence. *Knowledge and Information Systems, 33*(3), 523-547.

Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2016). *Opinion mining and sentiment polarity on twitter and correlation between events and sentiment.* Paper presented at the Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on.

Bell, D., Koulouri, T., Lauria, S., Macredie, R. D., & Sutton, J. (2014). Microblogging as a mechanism for human–robot interaction. *Knowledge-Based Systems, 69*, 64-77.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45-59.

Bing, L. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*: Cambridge University Press.

Blum, C., & Li, X. (2008). Swarm intelligence in optimization *Swarm Intelligence* (pp. 43-85): Springer.

Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems, 69*, 86-99.

Cambria, E., Rajagopal, D., Olsher, D., & Das, D. (2013). Big social data analysis. *Big data computing, 13*, 401-414.

Campbell, W. M., Dagli, C. K., & Weinstein, C. J. (2013). Social network analysis with content and graphs. *Lincoln Laboratory Journal, 20*(1), 61-81.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems, 51*, 70-82.

Ceran, B., Kedia, N., Corman, S. R., & Davulcu, H. (2015). *Story detection using generalized concepts and relations.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on.

Chen, S., Huang, Y., & Huang, W. (2016). *Big data analytics on aviation social media: The case of china southern airlines on sina weibo.* Paper presented at the Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on.

Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies, 7*(3), 246-259.

Cheng, Q., Li, T. M., Kwok, C.-L., Zhu, T., & Yip, P. S. (2017). Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *Journal of medical Internet research, 19*(7).

Cunha, J., Silva, C., & Antunes, M. (2015). Health twitter big bata management with hadoop framework. *Procedia Computer Science, 64*, 425-431.

Cybenko, G. (2017). *Parallel Computing for Machine Learning in Social Network Analysis.* Paper presented at the Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017 IEEE International.

Das, M., Elsner, M., Nandi, A., & Ramnath, R. (2015). *TopChurn: Maximum Entropy Churn Prediction Using Topic Models Over Heterogeneous Signals.* Paper presented at the Proceedings of the 24th International Conference on World Wide Web.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on, 6*(2), 182-197.

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services: Elsevier.

Do, N., Rahayu, W., & Torabi, T. (2016). A query expansion approach for social media data extraction. *International Journal of Web and Grid Services, 12*(4), 418-441.

Dong, X., Mavroeidis, D., Calabrese, F., & Frossard, P. (2015). Multiscale event detection in social media. *Data Mining and Knowledge Discovery, 29*(5), 1374-1405.

Elkaseh, A. M., Wong, K. W., & Fung, C. C. (2016). Perceived ease of use and perceived usefulness of social media for e-learning in Libyan higher education: A structural equation modeling analysis. *International Journal of Information and Education Technology, 6*(3), 192.

Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Engelbrecht, A. P. (2006). *Fundamentals of computational swarm intelligence*: John Wiley & Sons.

Erkens, M., Daems, O., & Hoppe, H. U. (2014). *Artifact Analysis around Video Creation in Collaborative STEM Learning Scenarios.* Paper presented at the Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on.

Evans, P. (2017). Generation and Analysis of a Social Network: Hamlet.

Fan, Z., Bi, D., He, L., Shiping, M., Gao, S., & Li, C. (2017). Low-level structure feature extraction for image processing via stacked sparse denoising autoencoder. *Neurocomputing, 243*, 12-20.

Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., & Quwaider, M. (2014). *Cross-lingual short-text document classification for facebook comments.* Paper presented at the Future Internet of Things and Cloud (FiCloud), 2014 International Conference on.

Fast, E., Chen, B., & Bernstein, M. S. (2016). *Empath: Understanding topic signals in large-scale text.* Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Fu, X., & Shen, Y. (2014). Study of collective user behaviour in Twitter: a fuzzy approach. *Neural Computing and Applications, 25*(7-8), 1603-1614.

Fulcher, J. (2008). Computational intelligence: an introduction *Computational intelligence: a compendium* (pp. 3-78): Springer.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR), 46*(4), 44.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Gao, B., Berendt, B., & Vanschoren, J. (2015). *Who is more positive in private? Analyzing sentiment differences across privacy levels and demographic factors in Facebook chats and posts.* Paper presented at the Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on.

Ghiassi, M., Zimbra, D., & Lee, S. (2016). Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. *Journal of Management Information Systems, 33*(4), 1034-1058.

Gilbert, C. H. E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text.* Paper presented at the Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf.

Gilbert, E., & Karahalios, K. (2009). *Predicting tie strength with social media.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Grinberg, N., Dow, P. A., Adamic, L. A., & Naaman, M. (2016). *Changes in engagement before and after posting to facebook.* Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Hajeer, M., Dasgupta, D., Semenov, A., & Veijalainen, J. (2014). *Distributed evolutionary approach to data clustering and modeling.* Paper presented at the Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on.

Harzing, A.-W. (2007). Publish or perish.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems, 47*, 98-115.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., . . . Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis, 35*, 18-31.

Hoeber, O., Hoeber, L., El Meseery, M., Odoh, K., & Gopi, R. (2016). Visual Twitter Analytics (Vista) Temporally changing sentiment and the discovery of emergent themes within sport event tweets. *Online Information Review, 40*(1), 25-41.

Hong, L., Dan, O., & Davison, B. D. (2011). *Predicting popular messages in twitter.* Paper presented at the Proceedings of the 20th international conference companion on World wide web.

Hu, W., Wang, H., Qiu, Z., Nie, C., Yan, L., & Du, B. (2017). An event detection method for social networks based on hybrid link prediction and quantum swarm intelligent. *World Wide Web, 20*(4), 775-795.

Huang, Y., Dong, H., Yesha, Y., & Zhou, S. (2014). *A scalable system for community discovery in twitter during hurricane sandy.* Paper presented at the Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems, 59*, 244-255.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems, 51*, 35-47.

Ismail, H. M., Zaki, N., & Belkhouche, B. (2016). *Using Custom Fuzzy Thesaurus to Incorporate Semantic and Reduce Data Sparsity for Twitter Sentiment Analysis.* Paper presented at the Soft Computing & Machine Intelligence (ISCMI), 2016 3rd International Conference on.

Ito, J., Song, J., Toda, H., Koike, Y., & Oyama, S. (2015). *Assessment of tweet credibility with LDA features.* Paper presented at the Proceedings of the 24th International Conference on World Wide Web.

Jaidka, K., Ramachandran, K., Gupta, P., & Rustagi, S. (2016). *Socialstories: Segmenting stories within trending twitter topics.* Paper presented at the Proceedings of the 3rd IKDD Conference on Data Science, 2016.

Jansson, P., & Liu, S. (2017). *Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media.* Paper presented at the Proceedings of the 3rd Workshop on Noisy User-generated Text.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.

Katal, A., Wazid, M., & Goudar, R. (2013). *Big data: issues, challenges, tools and good practices.* Paper presented at the Contemporary Computing (IC3), 2013 Sixth International Conference on.

Kennedy, J. (2006). Swarm intelligence *Handbook of nature-inspired and innovative computing* (pp. 187-219): Springer.

Kim, J., & Hastak, M. (2018). Social network analysis. *International Journal of Information Management: The Journal for Information Professionals, 38*(1), 86-96.

Kim, M., & Leskovec, J. (2012). Latent multi-group membership graph model. *arXiv preprint arXiv:1205.4546*.

Kim, W., Won, J. H., Park, S., & Kang, J. (2015). Demand forecasting models for medicines through wireless sensor networks data and topic trend analysis. *International Journal of Distributed Sensor Networks, 11*(9), 907169.

Kowalczyk, M., & Buxmann, P. (2015). An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study. *Decision Support Systems, 80*, 1-13.

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management, 34*(3), 387-394.

L'Heureux, A., Grolinger, K., ElYamany, H. F., & Capretz, M. (2017). Machine Learning with Big Data: Challenges and Approaches. *IEEE Access*.

Li, J., Rao, Y., Jin, F., Chen, H., & Xiang, X. (2016). Multi-label maximum entropy model for social emotion classification over short text. *Neurocomputing, 210*, 247-256.

Lim, S., Tucker, C. S., & Kumara, S. (2017). An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of biomedical informatics, 66*, 82-94.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*: Cambridge University Press.

Liu, W., Luo, X., Gong, Z., Xuan, J., Kou, N. M., & Xu, Z. (2016). Discovering the core semantics of event from social media. *Future Generation Computer Systems, 64*, 175-185.

Liu, W., Luo, X., Xuan, J., Xu, Z., & Jiang, D. (2016). Cognitive memory-inspired sentence ordering model. *Knowledge-Based Systems, 104*, 1-13.

Lohokare, J., Dani, R., & Sontakke, S. (2017). *Automated data collection for credit score calculation based on financial transactions and social media.* Paper presented at the Emerging Trends & Innovation in ICT (ICEI), 2017 International Conference on.

Lovinger, J., & Valova, I. (2015). *Scrubbing the Web for Association Rules: An Application in Predictive Text.* Paper presented at the Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on.

Lu, Y., Zhang, P., Cao, Y., Hu, Y., & Guo, L. (2014). On the frequency distribution of retweets. *Procedia Computer Science, 31*, 747-753.

Lv, Z., Shen, F., Zhao, J., & Zhu, T. (2016). *A swarm intelligence algorithm inspired by twitter.* Paper presented at the International Conference on Neural Information Processing.

Lyu, K., & Kim, H. (2016). Sentiment analysis using word polarity of social media. *Wireless Personal Communications, 89*(3), 941-958.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999): MIT Press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities, 2*, 460-475.

McCaig, D., Bhatia, S., Elliott, M. T., Walasek, L., & Meyer, C. (2018). Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities. *International Journal of Eating Disorders*.

Mukkamala, R. R., Hussain, A., & Vatrapu, R. (2014a). *Fuzzy-set based sentiment analysis of big social data.* Paper presented at the Enterprise Distributed Object Computing Conference (EDOC), 2014 IEEE 18th International.

Mukkamala, R. R., Hussain, A., & Vatrapu, R. (2014b). *Towards a set theoretical approach to big data analytics.* Paper presented at the Big Data (BigData Congress), 2014 IEEE International Congress on.

Nasukawa, T., & Yi, J. (2003). *Sentiment analysis: Capturing favorability using natural language processing.* Paper presented at the Proceedings of the 2nd international conference on Knowledge capture.

Nguyen, D. T., Joty, S., Imran, M., Sajjad, H., & Mitra, P. (2016). Applications of Online Deep Learning for Crisis Response Using Social Media Information. *arXiv preprint arXiv:1610.01030*.

Nguyen, T. H., & Shirai, K. (2015). *Topic modeling based sentiment analysis on social media for stock market prediction.* Paper presented at the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Nouh, M., & Nurse, J. R. (2015). *Identifying key-players in online activist groups on the facebook social network.* Paper presented at the Data Mining Workshop (ICDMW), 2015 IEEE International Conference on.

Ohbe, T., Ozono, T., & Shintani, T. (2017). *A sentiment polarity classifier for regional event reputation analysis.* Paper presented at the Proceedings of the International Conference on Web Intelligence.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1-135.

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing, 174*, 50-59.

Poria, S., Cambria, E., Hussain, A., & Huang, G.-B. (2015). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks, 63*, 104-116.

Povoda, L., Burget, R., Dutta, M. K., & Sengar, N. (2017). *Genetic optimization of big data sentiment analysis.* Paper presented at the Signal Processing and Integrated Networks (SPIN), 2017 4th International Conference on.

Qian, S., Zhang, T., Hong, R., & Xu, C. (2015). *Cross-domain collaborative learning in social multimedia.* Paper presented at the Proceedings of the 23rd ACM international conference on Multimedia.

Qian, S., Zhang, T., & Xu, C. (2014). *Boosted multi-modal supervised latent Dirichlet allocation for social event classification.* Paper presented at the Pattern Recognition (ICPR), 2014 22nd International Conference on.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing, 2016*(1), 67.

Reddick, C. G., Chatfield, A. T., & Ojo, A. (2017). A social media text analytics framework for double-loop learning for citizen-centric public services: A case study of a local government Facebook use. *Government Information Quarterly, 34*(1), 110-125.

Reuter, T., & Cimiano, P. (2012). *Event-based classification of social media streams.* Paper presented at the Proceedings of the 2nd ACM International Conference on Multimedia Retrieval.

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation, 47*(1), 239-268.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J, 2*(1), 127-133.

Samariya, D., Matariya, A., Raval, D., Babu, L. D., Raj, E. D., & Vekariya, B. (2016). *A Hybrid Approach for Big Data Analysis of Cricket Fan Sentiments in Twitter.* Paper presented at the Proceedings of International Conference on ICT for Sustainable Development.

Santos, M. C., Meira, W., Guedes, D., & Almeida, V. F. (2016). *Faster: a low overhead framework for massive data analysis.* Paper presented at the Cluster, Cloud and Grid Computing (CCGrid), 2016 16th IEEE/ACM International Symposium on.

Sato, S., Yonezawa, T., Nakazawa, J., Kawasaki, S., Ohta, K., Inamura, H., & Tokuda, H. (2016). *City happenings into wikipedia category: Classifying urban events by combining analyses of location-based social networks and wikipedia.* Paper presented at the Proceedings of the Second International Conference on IoT in Urban Space.

Schalkoff, R. J. (1997). *Artificial neural networks* (Vol. 1): McGraw-Hill New York.

Schulz, S., Pauw, G. D., Clercq, O. D., Desmet, B., Hoste, V., Daelemans, W., & Macken, L. (2016). Multimodular text normalization of dutch user-generated content. *ACM Transactions on Intelligent Systems and Technology (TIST), 7*(4), 61.

Serrat, O. (2017). Social network analysis *Knowledge solutions* (pp. 39-43): Springer.

Shanthi, C., & Pappa, N. (2017). An artificial intelligence based improved classification of two-phase flow patterns with feature extracted from acquired images. *ISA transactions, 68*, 425-432.

Simpao, A. F., Ahumada, L. M., Gálvez, J. A., & Rehman, M. A. (2014). A review of analytics and clinical informatics in health care. *Journal of medical systems, 38*(4), 45.

Song, S.-k., Kim, D. J., Hwang, M., Kim, J., Jeong, D.-H., Lee, S., . . . Sung, W. (2013). *Prescriptive analytics system for improving research power.* Paper presented at the Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on.

Spagnuolo, M., Maggi, F., & Zanero, S. (2014). *Bitiodine: Extracting intelligence from the bitcoin network.* Paper presented at the International Conference on Financial Cryptography and Data Security.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156-168.

Straton, N., Mukkamala, R. R., & Vatrapu, R. (2017). *Big social data analytics for public health: Predicting facebook post performance using artificial neural networks and deep learning.* Paper presented at the Big Data (BigData Congress), 2017 IEEE International Congress on.

Syeda, K. N., Shirazi, S. N. U. H., Naqvi, S. A. A., Parkinson, H. J., & Bamford, G. (2017). Big Data and Natural Language Processing for Analysing Railway Safety.

Takagi, H. (2001). Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE, 89*(9), 1275-1296.

Tan, Y., Mönch, L., & Fowler, J. W. (2014). *A decomposition heuristic for a two-machine flow shop with batch processing.* Paper presented at the Proceedings of the 2014 Winter Simulation Conference.

Tang, J., & Liu, H. (2012). *Unsupervised feature selection for linked social media data.* Paper presented at the Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

Thomaz, G. M., Biz, A. A., Bettoni, E. M., Mendes-Filho, L., & Buhalis, D. (2017). Content mining framework in social media: A FIFA world cup 2014 case analysis. *Information & Management, 54*(6), 786-801.

Trice, M. (2015). *Putting GamerGate in context: how group documentation informs social media activity.* Paper presented at the Proceedings of the 33rd Annual International Conference on the Design of Communication.

Tse, Y. K., Zhang, M., Doherty, B., Chappell, P., & Garnett, P. (2016). Insight from the horsemeat scandal: Exploring the consumers' opinion of tweets toward Tesco. *Industrial Management & Data Systems, 116*(6), 1178-1200.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234-246.

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics, 176*, 98-110.

Wlodarczak, P., Soar, J., & Ally, M. (2014). Big Data analytics of Social Media.

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering, 26*(1), 97-107.

Xu, Z., Liu, Y., Zhang, H., Luo, X., Mei, L., & Hu, C. (2017). Building the multi-modal storytelling of urban emergency events based on crowdsensing of social media analytics. *Mobile Networks and Applications, 22*(2), 218-227.

Yeon, H., & Jang, Y. (2015). *Predictive visual analytics using topic composition.* Paper presented at the Proceedings of the 8th international symposium on visual information communication and interaction.

Yu, S., Liu, M., Dou, W., Liu, X., & Zhou, S. (2017). Networking for big data: A survey. *IEEE Communications Surveys & Tutorials, 19*(1), 531-549.

Zhang, A. X., & Counts, S. (2016). *Gender and Ideology in the Spread of Anti-Abortion Policy.* Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.

Zhang, G., Sun, Y., Xu, M., & Bie, R. (2014). Weibo clustering: a new approach utilizing users' reposting data in social networking services. *Computer Science and Information Systems, 11*(3), 1157-1172.

Zhang, H., Sun, Z., Liu, Z., Xu, C., & Wang, L. (2015). *Dart: A geographic information system on hadoop.* Paper presented at the Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on.

Zhang, J., Ahlbrand, B., Malik, A., Chae, J., Min, Z., Ko, S., & Ebert, D. S. (2016). *A visual analytics framework for microblog data analysis at multiple scales of aggregation.* Paper presented at the Computer Graphics Forum.

Zitzler, E., Laumanns, M., Thiele, L., Zitzler, E., Zitzler, E., Thiele, L., & Thiele, L. (2001). SPEA2: Improving the strength Pareto evolutionary algorithm: Eidgenössische Technische Hochschule Zürich (ETH), Institut für Technische Informatik und Kommunikationsnetze (TIK) Zürich, Switzerland.

Acknowledgment

This work is supported by the UMRG Programme-AET (Innovative Technology (ITRC)), at the University of Malaya under grant RP0291-14AET and UMRG Programme - HNE (Humanities & Ethics), at the University of Malaya under grant RP044D-17HNE.

We review recent work on big social media analytics

We provide classification of big data analytics on social media

Recent advance in machine learning algorithms are discussed

We highlight some of the research challenges