Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Telematics and Informatics 57 (2021) 101517

Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele

Big data analytics meets social media: A systematic review of techniques, open issues, and future directions

Sepideh Bazzaz Abkenar [a], Mostafa Haghi Kashani [b], Ebrahim Mahdipour [a,*], Seyed Mahdi Jameii [b]

[a] *Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran*
[b] *Department of Computer Engineering, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran*

ARTICLE INFO

Keywords:
Social networks
Big data
Content analysis
Sentiment analysis
Systematic literature review

ABSTRACT

Social Networking Services (SNSs) connect people worldwide, where they communicate through sharing contents, photos, videos, posting their first-hand opinions, comments, and following their friends. Social networks are characterized by velocity, volume, value, variety, and veracity, the 5 V's of big data. Hence, big data analytic techniques and frameworks are commonly exploited in Social Network Analysis (SNA). By the ever-increasing growth of social networks, the analysis of social data, to describe and find communication patterns among users and understand their behaviors, has attracted much attention. In this paper, we demonstrate how big data analytics meets social media, and a comprehensive review is provided on big data analytic approaches in social networks to search published studies between 2013 and August 2020, with 74 identified papers. The findings of this paper are presented in terms of main journals/conferences, yearly distributions, and the distribution of studies among publishers. Furthermore, the big data analytic approaches are classified into two main categories: Content-oriented approaches and network-oriented approaches. The main ideas, evaluation parameters, tools, evaluation methods, advantages, and disadvantages are also discussed in detail. Finally, the open challenges and future directions that are worth further investigating are discussed.

1. Introduction

Social networking services (or social networking sites) are online platforms distributed across various computers over long distances. Millions of people all around the world use SNSs to upload photos, videos, update their current status, and post daily comments (Arora et al., 2019; Lai et al., 2020; Alalwan et al., 2017). They can join social networks in two ways: People may sign in by searching the network or may be invited by friends (Kumar et al., 2010). After being accepted by contacts, the inviter often invites the invitee's contacts, so the network expands in this way. The rapid growth of online social network relationship sites (e.g., Facebook, Myspace), media sharing networks (e.g., YouTube, Instagram), microblogging (e.g., Tumbler, Twitter), have encouraged researchers to investigate the published contents and analyse users' behaviors (Feng et al., 2018; Heidemann et al., 2012). A social network refers to structures among people or other social entities with the edges related to their associations (Busalim, 2016). In this structure, nodes are considered as people (or things) in the network, and interactions are expressed via the edges or links among them. A social network

* Corresponding author.
E-mail addresses: sepideh.bazzaz@srbiau.ac.ir (S. Bazzaz Abkenar), mh.kashani@qodsiau.ac.ir (M. Haghi Kashani), mahdipour@srbiau.ac.ir (E. Mahdipour), jamei@qodsiau.ac.ir (S.M. Jameii).

https://doi.org/10.1016/j.tele.2020.101517
Received 22 June 2020; Received in revised form 18 September 2020; Accepted 7 October 2020
Available online 14 October 2020
0736-5853/© 2020 Elsevier Ltd. All rights reserved.

originated in mathematical graph theory, which is defined as a graph, G= (V, E), in which V is a set of vertices or nodes that refers to people or objects, and E denotes a set of edges or ties indicating the relationships that connect the respective people (Bello-Orgaz et al., 2016).

Traditionally, data about users' interests and behaviors were collected by questionnaires. While this is still a prominent way in social science, the emergence and popularity of social networks have allowed us to collect data regarding users' behaviors in an unprecedented way, where we collect social data directly from users' social platform accounts (Jamali and Abolhassani, 2006). By collecting data from Online Social Networks (OSNs) and analyzing them, researchers can study a different aspect of users' behaviors and get valuable information (Martinez-Rojas et al., 2018; Cetto et al., 2018; Go and You, 2016). Now social science researchers send relevant queries to online social networks with the Application Programming Interface (API) to extract a large amount of users' data (Manovich, 2011). Further, most popular social networks provide API that allows researchers to gather and assess data from a given social media service (Lomborg and Bechmann, 2014). When the data is not accessible through API, researchers develop web crawlers to crawl OSN website, collect and extract data by using HTTP requests, manipulating, and responding to them (Abdesslem et al., 2012).

Thus, SNA is a scientific approach to extract data and analyse the structural characteristics of networks both quantitatively and qualitatively. In other words, the challenge of SNA is to study and extract the relationships among individuals, different organizations, and communities, which is essential for managing and reducing the complexities of social networks (Otte and Rousseau, 2002). Social network is an efficient way to collaborate and share knowledge among the core groups of the organization, research, and development units (Cross et al., 2002; Parveen et al., 2015). In this respect, the emergence of new social networks and an increasing number of social media users led to the explosion of user-generated contents (UGCs). Thus, it is crucial to know what this big data is and what insight can be gained from it (Boyd and Crawford, 2012). Big data refers to a massive volume and complicated amount of data that traditional tools are not able to manage and process effectively (Katal et al., 2013; Terrazas et al., 2019; Canito et al., 2018). Big data is different from "a large dataset" by the fact that the former is complex and has unique attributes, while the latter is a dataset with many records (di Bella et al., 2018).

In order to find out the role and influence of big data in social networks, the features of big data are described by using 5 V's, volume, velocity, variety, veracity, and value (Hadi et al., 2018). Volume means a vast amount of data that can be produced every second (Gandomi and Haider, 2015). Velocity stands for rapid generation of data, often referred to as streaming data (Kitchin, 2014). Variety represents various types of data, including structured, unstructured, and semi-structured data like images, videos, and texts (Sagiroglu and Sinanc, 2013; Pei et al., 2018). Veracity deals with the truthfulness of the data analysed and the accuracy behind any information (Bello-Orgaz et al., 2016). The value refers to the valuable information extracted for business and real values of the data (Peng et al., 2017). All these five features are available on social networks, so the most important application of big data is in the field of social media, which refers to *big social data* (or *social big data)*; the data are obtained from social networks. Big data technologies have produced new and exciting challenges in social networks (Duan et al., 2019).

Till today, with our observation and scrutiny, some surveys and Systematic Literature Reviews (SLRs) were performed on social big data analytics, but no comprehensive SLR has been written on social big data analytics that complicates the identification and assessment of the existing approaches, challenges, and gaps precisely. Moreover, due to the importance of big data analytics in social networks, this study aims at providing a systematic and comprehensive review to identify the challenges, potential future directions, merits, and demerits of this field. On the other hand, the association between SNA and big data analytic approaches is shown in particular and a research plan is investigated. An SLR presents a comprehensive review of state-of-the-art to reveal existing methods, challenges, and potential future research directions for research communities (Brereton et al., 2007). We conduct this SLR with the intention of *identifying, classifying, comparing social big data analytic approaches, evaluating the methods of existing papers systematically, and offering a reasonable taxonomy*. Additionally, to attain this intension and to answer the following research questions, this methodological review is conducted:

- Q1: What are the existing big data analytic approaches applied in social networks?
- Q2: What parameters do the researchers employ to evaluate the big data analytics in social networks?
- Q3: What are the tools used in social network analysis and big data areas?
- Q4: What are the social big data analysis applications in the studied papers?
- Q5: What are the datasets and case studies used in social big data analysis?
- Q6: What evaluation methods are applied to measure the big data analytic approaches in social networks?
- Q7: What are the challenges and future perspectives of big data analytic approaches in social networks?

We followed the guidelines in (Brereton et al., 2007; Kitchenham and Charters, 2007; Jamshidi et al., 2013; Jatoth et al., 2015) with the intension of exploring systematically, categorizing available social big data analytic approaches, and presenting a precise comparison analysis of approaches along with their potential challenges and limitations. This SLR presents a systematic review of the current studies on big data analytic approaches in social networks. For this purpose, 74 papers are chosen and compared to introduce a scientific taxonomy for the classification of big data analytic approaches in social networks. We summarize available methods, main ideas, applied tools, advantages, disadvantages, and evaluation parameters, and then provide statistical and analytical reports on them. Furthermore, this review identifies the motivation for presenting an SLR, outlines an abreast list of the primary challenges and open issues, and defines the significant areas where future research can improve the methods in the selected papers.

The remainder of this SLR is organized as can be seen in Fig. 1. Section 2 discusses some related works and motivation. The research questions, the details of the selection process, and the research methodology are documented in Section 3. Following, Section 4 provides a classification and a detailed study of the selected papers and demonstrates their main ideas, advantages, disadvantages,

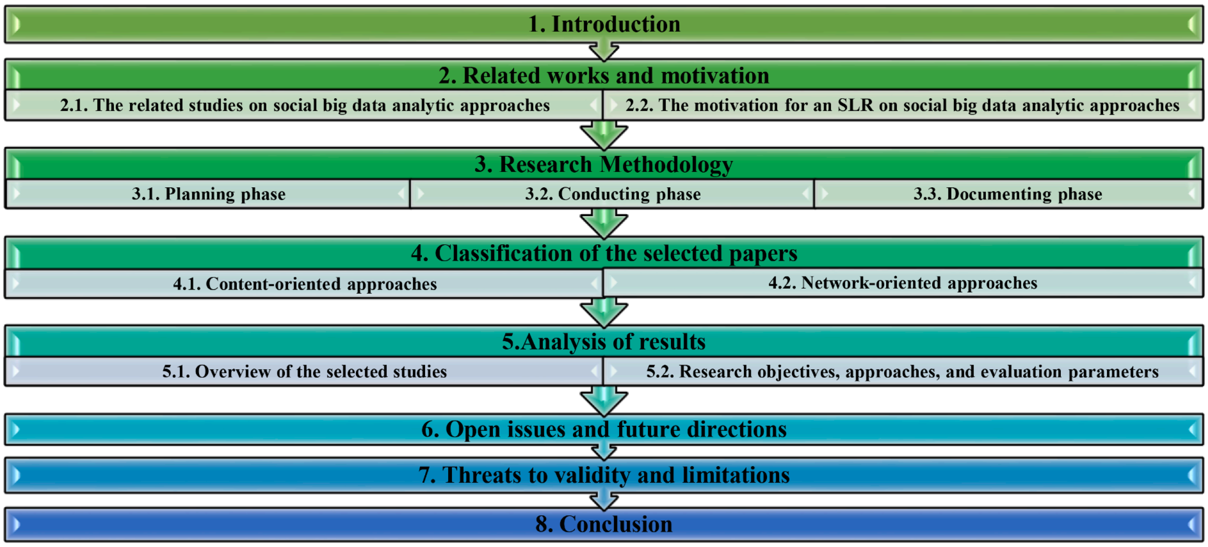
1. Introduction
2. Related works and motivation
2.1. The related studies on social big data analytic approaches
2.2. The motivation for an SLR on social big data analytic approaches
3. Research Methodology
3.1. Planning phase
3.2. Conducting phase
3.3. Documenting phase
4. Classification of the selected papers
4.1. Content-oriented approaches
4.2. Network-oriented approaches
5.Analysis of results
5.1. Overview of the selected studies
5.2. Research objectives, approaches, and evaluation parameters
6. Open issues and future directions
7. Threats to validity and limitations
8. Conclusion

Fig. 1. The structure of this SLR.

evaluation methods, tools, and evaluation parameters. Sections 5 and 6, respectively, disclose the analysis of the results, open issues, and future directions. Threats to validity and limitations are presented in Section 7. At last, the conclusion is explained in Section 8.

2. Related works and motivation

So far, there have been many reviews in the field of big data or social networks. However, the literature reviews conducted on this subject have some drawbacks. This section refers to several review studies that discussed social big data approaches.

2.1. *The related studies on social big data analytic approaches*

We explore the similarities and differences of the current reviews on this topic according to a systematic research, and the related works are summarized in surveys, and SLRs in Sections 2.1.1 and 2.1.2, respectively. Consequently, the weak points of these reviews are outlined in Section 2.1.3. In Table 1, a summary of the related works is illustrated in which such parameters as the main ideas, the review types, the paper selection processes, the taxonomies, open issues, evaluation parameters, applied tools, and the publication year of each study are represented.

2.1.1. *Surveys*

Yaqoob (2016) surveyed the possible applications of Information Fusion (IF) in social media. They also discussed social big data processing technologies, similarities, and differences based on relevant parameters. Moreover, the challenges of applying IF and future research directions were presented. The authors reviewed several potential applications of IF, such as advanced marketing, fraud detection, social context-based recommendation systems, and an advanced feasibility study was performed for new businesses and optimal decision making. Findings showed that applying fusion increases the accuracy, reliability, and confidence. However, business intelligence, integration, sharing, security, and data sharing were not touched in the paper. Besides, this research did not mention a systematic structure and the paper selection process was not clearly indicated.

Furthermore, di Bella et al. (2018) analysed the metadata for Scopus database papers in the field of big data in 1957–2017. The authors found that actual tendencies in academic big data literature were not enough in the building of real-time indicators considering this massive volume of productions. This study was written in a non-systematic manner and there was a gap among its discussions in big data quality measures, privacy, transparency, and big data diffusion. Moreover, recently published papers in the years 2018–2019 were not considered.

In another study, Ghani et al. (2018) provided a survey on social network analysis and classified the literature based on data sources, characteristics, computational intelligence, techniques of analysis, and the quality of features from the published papers between 2011 and 2017. The characteristics of big data analysis were summarized into descriptive, diagnostic, predictive, and prescriptive analytics. The authors classified the big data analytic techniques into modeling, sentiment analysis, SNA, and text mining. The papers were categorized according to approaches, techniques, and qualitative features by authors. Although, the paper selection process was not mentioned, they provided a comprehensive perspective of the big social media analytic research topics, and several challenges such as data quality, data locality, velocity, data availability, and natural language processing remained unaddressed.

Many other researchers perused several social big data papers such as (Bukovina, 2016) by reviewing technical analysis of social media to examine the behavior of capital markets, (Martin and Schuurman, 2019) by surveying social media data for qualitative

Table 1
Summary of the related works.

Type	Ref	Main idea	Pub. year	Paper selection process	Taxonomy	Open issue	Evaluation parameter	Applied tool	Covered year
Surveys	(Yaqoob, 2016)	Applications of information diffusion (IF) in social big data	2016	Not clear	No	Clear	Not presented	presented	Not mentioned
	(di Bella et al., 2018)	Big data and social indicators	2018	Not clear	No	Not clear	Not presented	Not presented	1957–2017
	(Ghani et al., 2018)	Big social media analytics	2018	Clear	Yes	Clear	Presented	Not presented	2011–2017
	(Bukovina, 2016)	Employing social big data analytics in the economic field	2016	Not clear	No	Clear	Not presented	Not presented	Not mentioned
	(Martin and Schuurman, 2019)	Analysis of social big data for GIScience	2019	Not clear	Yes	Not clear	Not presented	Not presented	Not mentioned
	(Arnaboldi et al., 2017)	Relationship between social media and big data and the accounting function	2017	Not clear	Yes	Clear	Not presented	Not presented	Not mentioned
	(Bello-Orgaz et al., 2016)	Social big data analysis	2016	Not clear	Yes	Clear	Not presented	presented	Not mentioned
	(Peng et al., 2016)	Influence analysis in social big data	2016	Not clear	No	Clear	Not presented	Not presented	Not mentioned
	(Guellil and Boukhalfa, 2015)	Social big data mining	2015	Not clear	Yes	Clear	Not presented	Not presented	2010–2015
	(Gole and Tidke, 2015)	Big data mining in social media	2015	Not clear	No	Not clear	Not presented	Presented	Not mentioned
	(Paul et al., 2017)	Big data analytics in social media	2017	Not clear	No	Not clear	Not presented	Presented	Not mentioned
SLRs	(Sebei et al., 2018)	Social media analytics within big data context	2018	Clear	Yes	Clear	Not presented	Presented	2008–2018
	(Al-Garadi, 2019)	Predicting cyber-attacks on social big data	2019	Clear	Yes	Clear	Not presented	Not presented	Not mentioned
	(Lerena et al., 2019)	Firm-level innovation based on social big data analysis	2019	Clear	Yes	Not clear	Not presented	Not presented	1970–2018
	This Study	Big data analytics in social networks	2020	Clear	Yes	Clear	Presented	Presented	2013–August 2020

geographic analysis, (Arnaboldi et al., 2017) by surveying the relationship between social big data analysis and the accounting function, (Bello-Orgaz et al., 2016) by reviewing the big data analytic algorithms in social media and their applications, (Peng et al., 2016) by conducting a survey to explore the architecture of influence analysis in social big data, and (Guellil and Boukhalfa, 2015; Gole and Tidke, 2015; Paul et al., 2017) by surveying big data mining in social media.

2.1.2. SLRs

Moreover, Sebei et al. (2018) presented an SLR by considering journal and conference papers published between the years 2008 and 2018 to provide a clear description of the social network analysis process applicable to big data technologies. In addition to suggesting solutions, the authors identified the challenges encountered during big data analysis. The social network analytic processes, challenges, solutions, and big data tools related to each step were studied, but the relevant parameters for comparing big data-related technologies were not specified.

Finally, other social big data SLRs are conducted such as (Al-Garadi, 2019) to detect cyber-attacks on social media via the aid of Machine Learning (ML) approaches, and (Lerena et al., 2019) by reviewing firm-level innovations based on text-mining and social network analysis.

2.1.3. Concluding remark

Considering the overviewed papers, some weaknesses have been noticed as described below:

- Some studies have not mentioned the periods of reviewed papers explicitly. In this paper, besides mentioning the scope of the study and the time range of articles, recently published articles have also been considered.
- The lack of a systematic construction in the related papers made the selection process unclear.
- Some papers have not been properly classified or have not presented any taxonomies. However, this paper not only provides a lucid and visual classification, but also defines a subclass for each of them.
- Some studies have not analysed the assessment parameters and evaluation tools. This SLR presents applied tools, evaluation parameters, and evaluation methods of the studied papers.
- Some of the related papers have not concentrated open issues explicitly, and future challenges have been enumerated briefly and implicitly. The presented literature is intended to highlight open issues well and precisely.

2.2. The motivation for an SLR on social big data analytic approaches

The need for an SLR is to *identify*, *classify*, and *compare the existing research reviews* on big data analytics in social networks. In order to show that a comprehensive SLR has not been already proposed, we searched Google Scholar with the following search string:

"big data"
[AND] Social
[AND]
(Review < OR > Overview < OR > Survey < OR > Challenges < OR > Study < OR> "open issues" <OR> "state-of-the-art" <OR > Trends)

According to the reasons mentioned in Section 2.1.3, and considering Table 1, most of the retrieved reviews were not conducted systematically, their paper selection processes were unclear, and they did not propose any lucid classification in their papers. To the best of our scrutiny, only three SLRs have been conducted on this topic (Sebei et al., 2018; Al-Garadi, 2019; Lerena et al., 2019) none of which has provided a complete systematic review to investigate SNA techniques, tools, strengths, weaknesses, open issues, evaluation parameters, and the application and critical role of big data in social networks. The two most similar efforts are in (Ghani et al., 2018), which is a survey not an SLR; It only covers journal papers between 2011 and 2017 and excludes conferences, and (Sebei et al., 2018), which is an SLR, covers the works between 2008 and 2018, but does not present evaluation parameters used in each studied paper. In (Al-Garadi, 2019), researchers only examined cyber-attacks and security issues in social big data, which differed from our paper, and the time range of studied papers was not specified. Additionally, open issues were not specified in (Lerena et al., 2019) and researchers in (Al-Garadi, 2019; Lerena et al., 2019) did not investigate the evaluation parameters and applied tools; therefore, writing an SLR that covers these weaknesses and highlights open issues and future research directions precisely is timely.

3. Research methodology

Researchers have conducted various studies on *social networks* and *big data*, their applications, and their challenges. In order to accomplish a comprehensive study of big data analytic approaches, this section presents an SLR method of big data analytic approaches in social networks. An SLR is a methodology to identify, classify, assess, and synthesize a comparative overview of the state-of-the-art in a specific subject (Brereton et al., 2007; Kitchenham et al., 2009). In contrast to other types of review papers, an SLR is a process of presenting a taxonomical review and performing a methodological analysis of the research literature to find the answers to problems and the given research questions related to specific research topics. The SLR has been used for the first time in medical fields (Aznoli and Navimipour, 2017) and can be conducted in any field of study for an accurate understanding, reducing bias, and identifying open issues and future directions (Rahimi et al., 2020; Haghi Kashani et al., 2020). Since most review articles on big data analytic approaches in social networks were written in unstructured procedures, the purpose of this paper is to provide a rigorous process of the methodological steps for researching the literature in this scope.

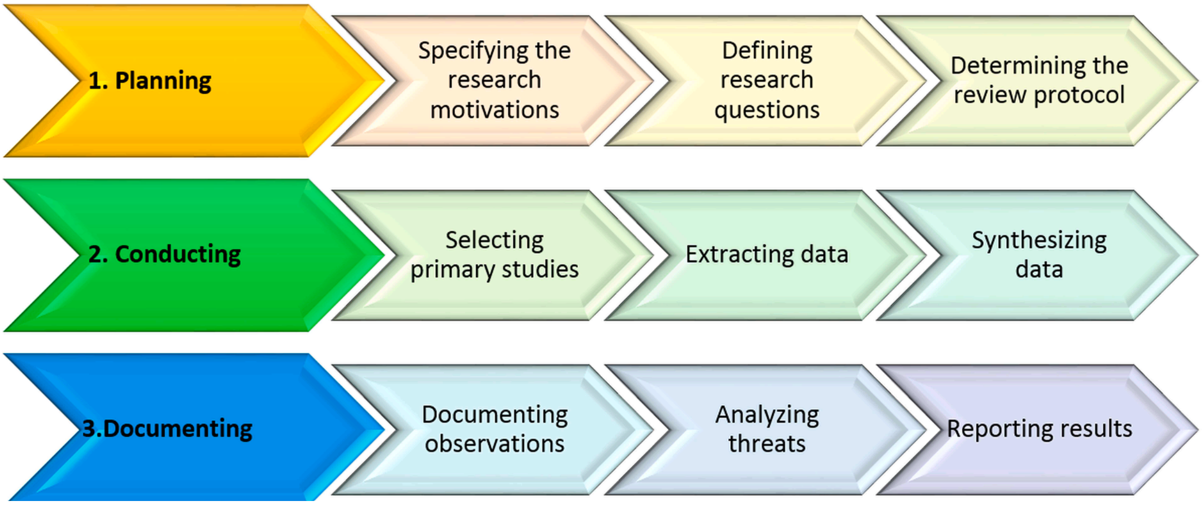

Fig. 2. Overview of research methodology.

In this systematic process, a three-phase guideline, namely *planning*, *conducting*, and *documenting* (Brereton et al., 2007) is adopted, as depicted in Fig. 2. The review is accompanied by an external evaluation of the outcome of each phase. We first identify the questions and the needs that are the motivation of this SLR in the planning phase. Then the articles in this subject are selected based on inclusion/exclusion criteria in the conducting phase. Ultimately, in the documenting phase, the observations are documented, and the results are analysed, compared, and visualized, which yields the answers to the research questions, then the final reports are represented. The three phases of the research methodology that are followed in this SLR are discussed below:

3.1. Planning phase

Planning begins with the determination of the research motivation for this SLR and finishes in a review protocol as follows:

Stage 1- *Specifying the research motivation.* According to the contribution of this SLR that is justified by comparing the available reviews explained in Section 2.2, the motivation is specified at the first stage.

Stage 2- *Defining research questions.* In the second stage, according to the motivation of this paper, the research questions are defined that assists the development and validation of the review protocol. The research questions are stated below. By finding the answers to the questions, available gaps on this subject can be found, which can facilitate reaching new ideas in documenting phase.

Q1: What are the existing big data analytic approaches applied in social networks?
Q2: What parameters do the researchers employ to evaluate the big data analytics in social networks?
Q3: What are the tools used in social network analysis and big data areas?
Q4: What are the social big data analysis applications in the studied papers?
Q5: What are the datasets and case studies used in social big data analysis?
Q6: What evaluation methods are applied to measure the big data analytic approaches in social networks?
Q7: What are the challenges and future perspectives of big data analytic approaches in social networks?

Stage 3- *Determining the review protocol.* According to the goals of this SLR, in the previous stage, the research questions and the review scope were identified to adjust search strings for literature extraction (Brereton et al., 2007). Moreover, a protocol was developed by following (Calero et al., 2013) and our previous experience with SLR (Haghi Kashani et al., 2020; Rahimi et al., 2020). To evaluate the defined protocol before its execution, we requested an external specialist for feedback, who was experienced in conducting SLRs in this era. His feedback was applied in the upgraded protocol. A pilot study (approximately 25%) of the included papers was performed to reduce the bias between researchers and to enhance the data extraction process. We also enhanced the review scope, search strategies, and inclusion/exclusion during the pilot stage.

3.2. Conducting phase

The second phase of the research methodology is conducting, starting with paper selection, and culminating in data extraction. This section aims to represent the process of searching and selecting papers conducted in the second phase of the SLR. The process of selecting papers consists of a three-step guideline as depicted in Fig. 3.

Table 2
Inclusion/Exclusion criteria.

	Criteria	Justification
Inclusion	Studies that focus on social big data analytics	Having a clear picture of big data analytic approaches in social networks
	Paper published online from 2013 to August 2020	The results of classical and fundamental literature on this subject have been mentioned in recent papers
Exclusion	Short papers that are less than six pages	These studies do not provide us with enough information to be used in our research.
	Surveys and review papers.	These studies do not offer any reasonable, significant, novel solutions, and information.
	Unjudged papers or papers that are not in English	Because of not trusting the quality of the unjudged papers and not having a possibility to probe non-English papers, these papers were excluded.
	Book chapters and theses	The result of book chapters or theses are mentioned in journal and conference papers

- ***First step.*** *The first step* of the research process was searching through Google Scholar[1] as the dominant search engine based on well-known academic publishers such as Springer[2], IEEE Explorer[3], ScienceDirect[4], SAGE[5], Taylor&Francis[6], Wiley[7], Emerald[8], ACM[9], and Inderscience[10] based on titles and keywords. The search strings were defined as follows:

("big data" <OR > Hadoop < OR > Spark < OR > Storm)
[AND]
("social media" <OR>"social networks" <OR> "social network" <OR > Twitter < OR > Facebook < OR > Instagram)

- ***Second step.*** At the end of the first step, 785 papers from journals, conferences, white papers, book chapters, and books were extracted. In *the second step*, the abstracts and conclusions of the papers were reviewed; then, papers published online from 2013 to August 2020 were chosen. Further, to obtain the most relevant papers, non-peer-reviewed and non-English papers, theses, review papers, short papers, and book chapters were excluded (as shown in Table 2). At this step, we found 246 papers covering social big data analytics.

- ***Third step.*** Finally, in *the third step*, the full texts of all selected papers were reviewed, and for further detailed analysis, 74 relevant papers were chosen, which could answer our research questions and fully describe the methods and challenges. Investigating 74 relevant papers assists us in proposing a classification on social big data analysis approaches in Section 4 and revealing the pros and cons of these approaches.

3.3. Documenting phase

As determined in Fig. 2, in documenting phase, after documenting the observations, threats to validity and limitations are explored which is presented in Section 7. Then the results are analysed, visualized, and reported in Section 5.

4. Classification of the selected papers

In this section, 74 chosen papers are explored to examine social big data analysis objectives, techniques, and innovations; a review of the advantages and disadvantages of each approach is also presented. A taxonomy of the related literature is given in this paper, and the pictorial description of the proposed taxonomy for the reviewed papers is shown in Fig. 4. Offering a taxonomy for social big data analysis is not a trivial and easy task. As researchers look at the problems in this area from various perspectives, each researcher performs this classification differently. By using this categorization, the reader can easily refer to each of these papers as a categorical reference. The selected papers use big data analytic techniques for analyzing social networks. These techniques are categorized into

[1] https://scholar.google.com
[2] https://link.springer.com
[3] https://ieeexplore.ieee.org
[4] https://www.sciencedirect.com
[5] https://online.sagepub.com
[6] https://www.tandfonline.com
[7] https://onlinelibrary.wiley.com
[8] https://www.emeraldinsight.com
[9] https://dl.acm.org
[10] https://www.inderscienceonline.com

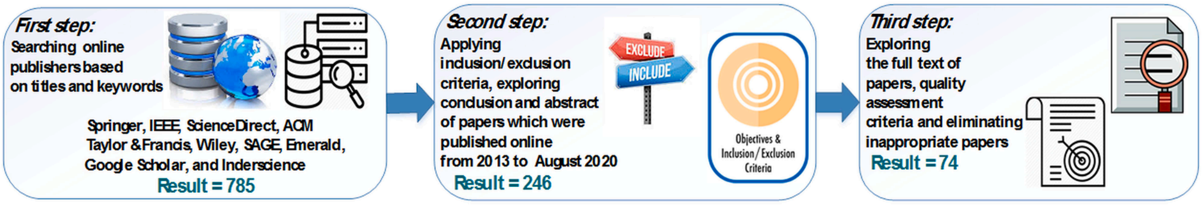
First step:
Searching online publishers based on titles and keywords
Springer, IEEE, ScienceDirect, ACM Taylor &Francis, Wiley, SAGE, Emerald, Google Scholar, and Inderscience
Result = 785
Second step:
Applying inclusion/exclusion criteria, exploring conclusion and abstract of papers which were published online from 2013 to August 2020
EXCLUDE
INCLUDE
Objectives & Inclusion/Exclusion Criteria
Result = 246
Third step:
Exploring the full text of papers, quality assessment criteria and eliminating inappropriate papers
Result = 74

Fig. 3. Paper selection process.

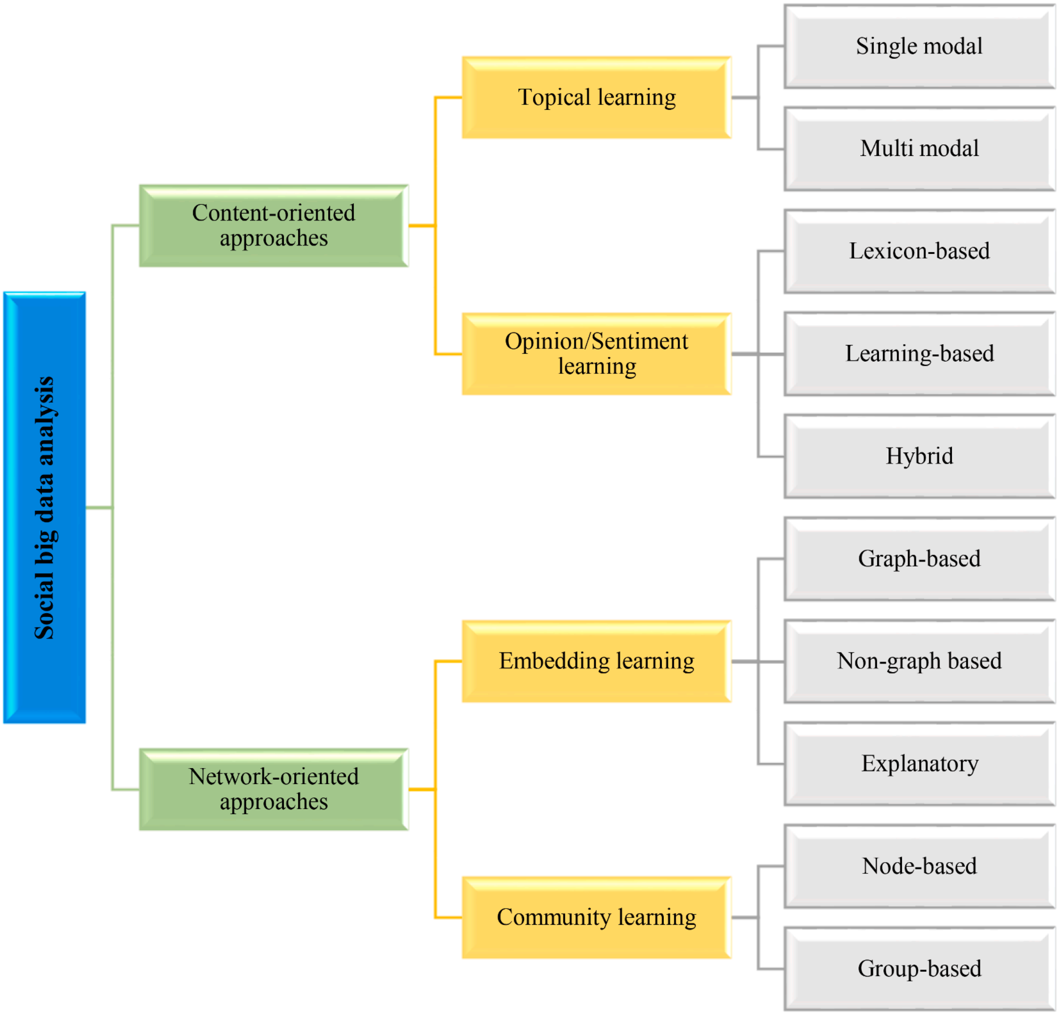

Fig. 4. Taxonomy of social big data analysis.

two major groups: Content-oriented approaches, and network-oriented approaches.

Content-oriented approaches are classified into two subgroups, namely topical learning and opinion/sentiment learning. Topical learning can be performed in a single modal or a multimodal approach. Opinion/sentiment learning can be carried out in lexicon-based, learning-based, or hybrid approaches. Further, network-oriented approaches are classified into two groups: Embedding learning and community learning. Embedding learning has graph-based, non-graph based, and explanatory models, while, community learning is node-based or group-based. The papers relevant to content-oriented approaches and network-oriented approaches are reviewed in Sections 4.1 and 4.2, respectively. In this study, the methods of big data analysis on social networks are examined and evaluated with a list of important evaluation parameters. Further, the definition associated with evaluation parameters of the reviewed papers, as well as their formulas, is presented in Appendix A.

4.1. Content-oriented approaches

Nowadays, with the explosion of data in social networks that provides the researchers with a different type of contents instead of the traditional books and libraries, it is essential to analyse this immense volume of data. In this paper, the selected papers with topical learning and opinion/sentiment learning are reviewed in Sections 4.1.1 and 4.1.2, respectively. In Sections 4.1.1 and 4.1.2, classification of techniques, the definition of methods, and the related papers are discussed.

4.1.1. Overview of the topical learning approaches

In content-oriented approaches, topical learning focuses on the communication contents of social networks, consisting of text mining, video content analysis, and image analysis. It is the process of analyzing various types of unstructured data, like images, audio and video files, or different types of text including word, PDF files, PowerPoint slides, posts of weblogs and social network sites, or semi-structured data such as XML, HTML, JSON, and CSV files with the purpose of uncovering underlying similarities and hidden associations and transforming them into structured data for further analysis. The topical learning may be either performed “single modal” or “multi-modal” in which a “single modal” collects and analyses one modality (text OR audio OR image OR video) whereas “multi-modal” analyse a combination of various types of datasets such as text, audio, image, and video. According to the reviewed papers, the comparison between the specification and evaluation parameters is illustrated in Tables 3 and 4. Table 3 summarizes the main ideas, advantages, disadvantages, evaluation methods, tools, and case studies along with their categories related to the papers in this approach. Table 4 presents a side by side comparison of the evaluation parameters in papers related to topical learning approaches.

In order to investigate the effects of social media on Eating Disorders (ED), Moessner et al. (2018) applied texts, linguistics, and lexical analysis with an unsupervised, bottom-up method to identify harmful posts. They did not investigate social media data in real-time, otherwise, the safety of ED-related communication could have been improved. Further, to execute the balance policies in the business application of social networks, Huo et al. (2018) presented a new logic Datalog. TS_u_Datalog was presented as the most appropriate logic Datalogs and a new programming language with both Active_U_Datalog and Distributed Temporal Logic (DTL) was introduced to implement contractual policies in a dynamic social media. The results of the time evaluation parameter of TS_u_Datalog could have been improved and used for blockchain systems, privacy-preserving of smartphones, and as a fault tolerance technique for wireless sensor networks.

To enhance health monitoring systems to detect infectious disease and to take preventive actions, Zadeh et al. (2019) presented a spatio-temporal platform to check out whether social posts could discover flu outbreaks in a particular area during the flu season. As some people do not activate their GPS or do not express their geographic locations in a social network profile, the geographic analysis cannot be done more deeply and accurately. More efficient ML techniques were needed to perform more in-depth analysis and to identify noise and unrelated social network posts. To recognize all repetitive and non-repetitive substring in passwords, Xylogiannopoulos et al. (2020) designed an efficient pattern detection system that can be embedded in social network platforms to generate a more robust and valid password. The results indicated that, contrary to common belief, long passwords are not safe, but passwords that are a combination of small/capital numbers and symbols are stronger than the others. This methodology did not have a limitation on the length and the type of characters. However, the proposed system could have been tested on other datasets, leading to different results.

In order to prevent the death caused by Adverse Drug Reactions (ADRs), Yang et al. (2015) used text classification to propose an automated framework to filter ADR related posts. A supervised learning method was applied to classify the extracted posts into positive/negative examples. The results of classification were used as an input to build an early warning system to prohibit future ADRs. Although the presented method generally outperformed in precision, recall, and F-measure, they did not extend their framework for various types of drugs. Furthermore, Cheung et al. (2015) presented a connection discovery system for follower/followee recommendations instead of user-generated tags and social graphs. They used Bag-of-Features Tagging (BoFT) to label user-shared images with BoFT labels, and a computer vision approach was employed to model the characteristics of user-shared images. In addition to the identification of user’s gender in the proposed system, the image classification performance was higher than K-mean, and there was no need to know K (the number of clusters in the clustering) in advance. However, the runtime of clustering and feature extraction was high. Subsequently, for more users and user-shared images, a big data system is required to manage and discover data.

Furthermore, to identify mental disorders in advance, Thorstad and Wolff (2019) scrutinized people’s every day mental and non-mental health topic posts on Reddit website. The outcome of the accuracy assessment indicated that people’s posts on clinical and non-clinical subreddits were highly and moderate predictive of mental disease, respectively. Also, it revealed that the predictions were more precise on recent past posts compared to distant past posts. The limitation was that posting a clinical post may not be a significant criterion for early diagnosis of psychological disorders, as some people may be affected by mental illnesses before posting. Besides, to identify vulnerabilities, Subroto and Apriyana (2019) offered an algorithmic model applying social media analytics and ML algorithms to protect cyber-attacks. Despite the highest accuracy of the model created by artificial neural networks, it was not scalable, having hardware limitations, and was tested only on a small sample of Twitter dataset, but the authors claimed that it did not affect the accuracy of the model.

Moreover, many other studies adopted clustering and ML algorithms in text mining and trending topics on big data of social platforms (Straton et al., 2017; Makaroğlu et al., 2019; Vakali et al., 2016; Aa et al., 2015). Also, researchers in (Singh and Kaur, 2019; Sachar and Khullar, 2017) proposed hybrid models by applying a metaheuristic approach to enhance the classification performance in the content analysis of social big data. Nowadays, as millions of users produce and share videos in various social media, Panarello et al. (2020) developed a framework for video transcoding processing in a short time. They applied Hadoop in their cloud federation framework to transcode videos to be compatible with sharing of users with different hardware/software devices. The evaluation results

Table 3
Reviewing and comparing papers with topical learning approaches.

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
Single modal	(Moessner et al., 2018)	Creating a linear network autocorrelation model	• Analyzing the text and network of an eating disorder forum	• Not able to investigate social media in real-time	Real test bed	MySQL database, RMySQL, R studio, R programming language	The proed forum on www.reddit.com
	(Huo et al., 2018)	Proposing a novel logic and flexible TS_u_Datalog	• Using a flexible datalog logic to execute contractual policies and to keep balance among users, advertisers, and social network providers	• Low privacy • Low reliability	Example application	Not mentioned	Not mentioned
	(Zadeh et al., 2019)	Presenting spatio-temporal big data analysis to detect real-time behavioral patterns during the flu season	• High scalability • High accuracy	• Lacking geographic information of all users in the social network	Real test bed	Hadoop,Big R released by IBM,Sqoop,Apache Flume	Twitter,Cerner HealthFacts data warehouse
	(Xylogiannopoulos et al., 2020)	Presenting a password creation and validation system for social media platforms	• High accuracy • High scalability • Response time • Low cost	• Unacceptable security levels according to the needs of the social network platform	Real test bed	C#, SQL Server 2014	LinkedIn
	(Yang et al., 2015)	Proposing a new early warning system for adverse drug reactions	• High recall • High F-measures • High accuracy • High precision • Low cost	• Not extending the proposed method for other types of drugs	Data sets	Not mentioned	the online health community, MedHelp
	(Cheung et al., 2015)	Proposing a recommendation system using big data of user-shared images in social media	• High accuracy • Able to discover connection and gender identification • Not requiring to know k (the number of clusters in clustering) in advance	• Having a high runtime of feature extraction and clustering for more users	Simulation	Matlab	Skyrock,Sina Weibo,Flickr
	(Thorstad and Wolff, 2019)	Presenting a multiclass classification to reveal mental disorders by investigating people's posts on Reddit website	• High accuracy • High precision • High recall • High F-measure	• Undermining the future mental disorder prediction claims in case of being affected before posting	Data sets	Python, Scikit-learn library	Reddit website
	(Subroto and Apriyana, 2019)	Presenting an algorithmic model employing social media analytics and statistical machine learning to predict cyber risks	• High accuracy • High precision • High recall	• Low scalability	Data sets	MySQL, Rweka package, RStudio (R Statistical software)	Twitter
	(Straton et al., 2017)	Applying artificial neural networks and deep learning to predict Facebook posts	• High accuracy • Low run time	• Low scalability	Data sets	Not mentioned	Facebook
	(Makaroğlu et al., 2019)	Analyzing Turkish news on Twitter with Apache Spark	• High scalability	• Low security	Data sets	Python,Apache Spark	Twitter
	(Vakali et al., 2016)	Presenting a framework for trend detection in social networks	• Low running time (high speed) • High accuracy • High recall	• Not applying new technologies	Real test bed	Hadoop,Apache Drill, Apache Storm	Twitter
	(Aa et al., 2015)		• High accuracy	• Low scalability	Data sets	Apache Giraph,Apache Hive	Facebook

(*continued on next page*)

Table 3 (*continued*)

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
		A novel face recognition framework in social networks based on ML	• High robustness				
	(Singh and Kaur, 2019)	Presenting a hybrid content-based cyberbullying detection model based on the metaheuristic approach in social networks	• High recall • High precision • High predictive performance in cyber-crime datasets	• Low scalability	Data sets	Python	Twitter, ASKfm, FormSpring
	(Sachar and Khullar, 2017)	Applying genetic algorithm in clustering social big data	• Low cost • Low execution time • High scalability • High accuracy	• Not applying another metaheuristic algorithm • Not evaluating other parameters	Data sets	Hadoop, Java, Mahout	Twitter
	(Panarello et al., 2020)	Offering a framework for analyzing the video transcoding based on cloud	• High scalability • Low processing time	• Not investigating the privacy and security of the presented framework	Real test bed	Hadoop, NoSQL, Amazon S$_3$(Amazon cloud storage provider),CLEVER (Cloud-Enabled Virtual Environment)	Not mentioned
	(Alomari et al., 2020)	Presenting a traffic event detection tool	• High accuracy • High precision • High recall • High F-measure	• Low scalability	Data sets	Apache Spark,MongoDB, Python	Twitter
Multimodal	(Zhou et al., 2016)	Introducing a private video recommendation system based on cloud and online learning	• High accuracy • High privacy-preserving level	• Low scalability	Simulation	Not mentioned	Sina microblog, Youku (video sharing site)
	(Feng et al., 2018)	Proposing a content-centric networking architecture based on Monte Carlo Tree Search	• High scalability • High accuracy • Low running time • High security • Low cost	• Low energy efficiency	Simulation	Not mentioned	Sina Weibo
	(Sahoo and Gupta, 2020)	Presenting a Facebook fake profile detection framework	• High recall • High accuracy • High precision • High F-measure • High ROC (AUC) • High Matthews's correlation coefficient • High TPR	• Not evaluating the responding time of the presented approach	Data sets	Weka	Facebook
	(Zhang et al., 2019)	Presenting a multi-modal microblog emotion analyzer based on deep learning	• High accuracy • High recall • High F-measure	• Not considering the personality and expression habits of a single user	Data sets	Not mentioned	Sina Weibo

Table 4
An overview of the evaluation parameters in papers with topical learning approaches.

Category	Ref.	Centrality Measures	Security	Accuracy	Precision	Recall	F-measure	Scalability	Time	Cost	ROC (AUC)	Specificity	Matthews correlation coefficient
Single modal	(Moessner et al., 2018)	✓											
	(Huo et al., 2018)								✓	✓			
	(Zadeh et al., 2019)			✓	✓	✓	✓	✓					
	(Xylogiannopoulos et al., 2020)			✓				✓	✓	✓			
	(Yang et al., 2015)			✓	✓	✓	✓			✓			
	(Cheung et al., 2015)			✓	✓	✓			✓				
	(Thorstad and Wolff, 2019)			✓	✓	✓	✓						
	(Subroto and Apriyana, 2019)			✓	✓	✓							
	(Straton et al., 2017)			✓					✓				
	(Makaroğlu et al., 2019)	✓						✓					
	(Vakali et al., 2016)			✓		✓			✓				
	(Aa et al., 2015)			✓									
	(Singh and Kaur, 2019)				✓	✓	✓				✓		
	(Sachar and Khullar, 2017)			✓			✓	✓	✓	✓			
	(Panarello et al., 2020)								✓				
	(Alomari et al., 2020)			✓	✓	✓	✓						
Multi modal	(Zhou et al., 2016)		✓	✓									
	(Feng et al., 2018)		✓	✓				✓	✓	✓			
	(Sahoo and Gupta, 2020)			✓	✓	✓	✓				✓	✓	✓
	(Zhang et al., 2019)			✓		✓	✓						

on real testbed demonstrated performance enhancement in terms of speed, scalability, and transcoding time, but security and privacy issues were neglected.

Alomari et al. (2020) developed a methodology based on text mining by using big data technologies for road traffic detection from Arabic tweets. The authors applied three machine learning algorithms, namely Logistic Regression, Support Vector Machine, and Naïve Bayes for classifying eight types of events. The evaluation results showed enhancement in text processing, leading to more accurate event detection with no prior knowledge about those events. However, this methodology could also be used to identify events other than road transportation. They did not focus on improving scalability and data management of the proposed method.

Zhou et al. (2016) proposed a private video recommendation system based on distributed online learning. Multimedia such as images, audios, and videos produced by users were sent and stored in remote and decentralized data centers. The user's context vectors were extracted by BOFT (bag-of-features tagging) and converted into distributed video service servers. At last, the recommended video was transferred to multimedia applications in online social networks. The evaluation results on real datasets in Sina microblog and Youku, a video sharing site (VSS) in China, achieved sublinear regret bound and established a trade-off between the performance loss and the privacy protection level. However, for simplicity, a small dataset was chosen in those social networks, so it suffered from low scalability. In another study, Feng et al. (2018) proposed a Content-Centric Networking (CCN) architecture based on the Monte Carlo Tree Search (MCTS) algorithm. Since the volume and variety of both users and contents are rapidly growing, the MCTS algorithm solved the accurate content push problem in big data. Their algorithm outperformed in the experimental results of push accuracy, scalability, and robustness of users' arrivals in Sina Weibo on an offline dataset. Although the proposed architecture could evaluate the performance in a real-world CCN-based social media, energy efficiency was neglected.

Sahoo and Gupta (2020) proposed a framework to distinguish fake profiles on Facebook. The authors applied various ML algorithms along with content analysis and account-based features to detect suspicious accounts from genuine ones. The evaluation results indicated that the presented framework gave the best outcome in terms of accuracy, precision, recall, F-measure, and Matthews's correlation coefficient, but they did not evaluate the responding time of the presented approach. Moreover, applying this approach on other platforms such as Twitter and Google + or adding an aggregator module for comparing various account features and their activities may lead to different results. Since various microblogs contain videos, emoticons, and pictures as well as texts, Zhang et al. (2019) proposed a multi-modal emotion analyzer based on deep learning. The authors applied a two-way Long and Short Term Memory network (LSTM) model to integrate contents and user's features. The offered model attained a higher accuracy, precision, and F-measure compared to previous models, but users' personalities were not considered and in the proposed model, user-based emotions could not be classified.

4.1.2. Overview of the opinion/sentiment learning approaches

In this section, the selected papers with opinion/sentiment learning approaches are reviewed. Opinion/sentiment learning approaches entail Natural Language Processing (NLP) to extract opinions from the text and classify the polarity of subjects into positive, negative, or neutral to determine what they are talking about and to identify the public group perception. With the help of sentiment analysis, opinions about products, services, brands, politics, or any topic that people care about are extracted. These data can be used in many applications like marketing analysis, product reviews and feedback, emotion detection, intent analysis, customer support and services, social media monitoring, and brand monitoring (Shirdastian et al., 2019).

By reviewing papers relevant to opinion/sentiment learning, we recognized three methods, namely lexicon-based, learning-based, and hybrid approaches, employed to extract and analyse opinion/sentiment in social media contents. In lexicon-based approaches, a set of predefined lexical wordlist, corpus, and dictionaries are used to extract subjectivity, the orientation, and the polarity of opinions and sentiments. Learning-based approaches utilize various ML algorithms (supervised or unsupervised) to classify text into positive or negative classes. Moreover, some of the reviewed papers combine both learning-based and lexicon-based approaches that mentioned hybrid approaches. Table 5 depicts a comparison of the selected papers with opinion/sentiment learning approaches. It includes main ideas, advantages, disadvantages, evaluation methods, tools, and case studies along with their categories. In some studies, the applied tools for analyzing and implementing approaches have not been mentioned. Table 6 shows the parameters used by papers relevant to opinion/sentiment learning approaches to evaluate the intended methods.

Kauffmann et al. (2019) offered a modular framework for qualitative interpretation of UGC by employing NLP techniques and applying cosine similarity measures to recognize fake reviews. Their Fake Review Detections Framework (FRDF) utilized NLP techniques to discover similarities between reviews and eliminate fake and unreliable reviews of a product. The major weakness was that FRDF set a threshold in the cosine similarity measure to detect fake reviews; other thresholds or other sentiment analysis tools, except for lexicon Afinn, may produce different outcomes. Furthermore, Jiang et al. (2017) suggested a method for performing sentiment computing of the news event in social big data. First, a Word Emotion Association Network (WEAN) was constructed to compute both word and text emotions at a specific time. After dividing emotions, a questionnaire was designed to collect ideas about the six-dimensional sentiment emotion of emoticons, and emoticons were used to calculate the emotions of each sentence. Second, based on WEAN, a word emotion computation algorithm was presented to get the primary word emotions. Then an emotional refinement algorithm was offered by employing the standard emotional thesaurus to improve the sentiments of news with high accuracy, but emotion distance and word emotion patterns were not considered into text sentiment computations.

Moreover, Dalla Valle and Kenett (2018) presented a new approach to integrate online review data with customer survey data. The sentiments of online users were calibrated with customer surveys by resampling and merging data via Bayesian networks in their method. This approach was used in various areas, and the data integration between online blogs and customer satisfaction led to enhancement in sentiment analysis. However, it did not consider methods for integrating vast data sources to enhance the accuracy of results. In addition, Jimenez-Marquez et al. (2019) presented a two-stage framework to analyse UGC in social media. The first stage,

Table 5
Reviewing and comparing papers with opinion/sentiment learning approaches.

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
Lexicon-based	(Kauffmann et al., 2019)	Presenting a fake review detection framework for sentiment analysis of social networks	• High accuracy • High recall • High F-measure • Modularity • High flexibility	• Not being evaluated on different lexicon except for lexicon Afinn • Not being tested on various cosine similarity thresholds	Data sets	R language,Set of NLP tools such as Stanford CoreNLP, OpenNR, Tidytext, Afinn sentiment lexicon	Amazon website
	(Jiang et al., 2017)	Introducing a sentiment computing method based on the social media big data	• High accuracy	• Not considering emotion distance and word emotion pattern into text sentiment computation	Data sets	Not mentioned	Sina microblog
Learning-based	(Dalla Valle and Kenett, 2018)	Introducing a data integration approach based on calibration	• High accuracy • Enhancing the information quality of the data analytic namely, data structure, data integration, temporal relevance, and chronology of data, and goal, and calibrating the sentiments of online blogs and a customer satisfaction survey	• Low accuracy in case of a large number of data sources	Data sets	GeNIe Software V 2.1, R package ROSE	San Francisco international airport passengers dataset, Skytrax dataset
	(Jimenez-Marquez et al., 2019)	Proposing a two-stage big data and ML framework to analyse social media content	• High accuracy • High precision • High recall • High flexibility	• Low scalability	Data sets	Spark, Python, MySQL, Natural Language Toolkit (NLK & Pandas package)	Tourism data from Yelp dataset (Yelp.com)
	(Zhu et al., 2020)	Analyzing the opinions of users on COVID-19 epidemic on microblog	• The results of the analysis can assist in managing and controlling public health	• The analysis was limited to the provincial level • Not considering the age and gender of users	Data sets	Python	Sina Weibo
	(Fan et al., 2020)	Presenting an ML model to analyse the tweets of English national team fans during 2018 FIFA world cup	• Sentiment analysis of tweets as well as emoji analysis • High accuracy	• Low F-measure • Low precision • Low recall • Low reliability due to multilingual and sarcastic tweets	Data sets	Python	Twitter
	(Shirdastian et al., 2019)	Offering a framework to explore brand validity sentiments	• High precision • High accuracy	• Not exploring the brand validity sentiments over time • Not excluding the sentiments created by bots	Data sets	Python	Twitter
	(Lee and Paik, 2017)	Presenting a real-time processing system to analyse stock market tweets	• High classification accuracy • High scalability • Real-time processing	• Not able to process complex data	Data sets	Apache Spark, Apache Kafka	Twitter
	(Sayed et al., 2020)	Presenting a metaheuristic approach in sentiment analysis of tweets	• High accuracy • Real-time analytics	• Not evaluating the accuracy of the proposed model by various evaluation parameters	Data sets	Apache Spark	Twitter
Hybrid	(van Dieijen et al., 2020)	Presenting a framework to examine the relationship between volatility in the stock markets and UGCs	• High accuracy	• Low scalability • Time-consuming • High cost	Data sets	Matlab	Twitter

(continued on next page)

Table 5 (*continued*)

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
	(Spruce et al., 2020)	Offering a new methodology for sentiment analysis to explore the impact of social sensing on weather events	• High accuracy	• Low scalability • Low reliability	Data sets	Python	Twitter
	(Um et al., 2013)	Introducing a distributed and parallel parsing system on the MapReduce framework	• High precision • High recall • High accuracy • Extracting information very fast • Reducing parsing time (Response time) • High scalability • High portability (Low cost)	• Not being tested on actual Twitter SNS data • Not analyzing specialized sentences as well as ordinary user's statements technically	Data sets	Hadoop, Java	KISTI, NDSL
	(Moise, 2016)	Analyzing tweets regarding vaccination sentiments and their trends in Twitter	• High scalability	• Not evaluating the time parameter	Real test bed	Hadoop, Mahout	Twitter
	(Hsu et al., 2017, 2017)	Mining tweets that contain reporting on drug side effects	• High F-measure • High speed of extraction and processing tweets • High accuracy • High scalability	• In case of tweets about multiple drugs that cause multiple side effects, a manual examination is needed to see which side effect corresponds to each drug	Data sets	Apache Spark's ML library (MLlib), Python	Twitter
	(Baltas et al., 2016)	Presenting a sentiment analysis framework by applying ML techniques	• High F-measure	• Low scalability • Not evaluating the time parameter	Data sets	Apache Spark's ML library (MLlib)	Twitter
	(Sun et al., 2018)	Designing a microblog abnormal emotion detection model based on the neural network and CNN-LSTM	• High accuracy • High recall • High F-measure • Detecting abnormal user's emotions on a microblog that can be used in public security monitoring • Dynamic and real-time detection	• Time performance (Improving the threshold selection which is a time-consuming process)	Data sets	Not mentioned	Sina Weibo
	(BalaAnand et al., 2019)	Presenting a mechanism to gather and to envision social media information for big data	• High accuracy • High scalability • High robustness • Low computational time • High precision • High recall • High F-measure	• Not being assessed on the cost of the proposed method	Data sets	Apache Flume, Hadoop, Java platform	Twitter,Facebook, Amazon dataset, Kaggle dataset
	(Rodrigues and Chiplunkar, 2019)	Proposing a topic classification and sentiment analysis framework	• High accuracy • High precision • High recall • High F-measure • Low execution time • Low responses time	• Low performance in case of sarcastic opinions and cross-lingual sentiment classification	Data sets	Hadoop platform, Apache Flume,Apache Hive	Twitter

Table 6
An overview of the evaluation parameters in papers with opinion/sentiment learning approaches.

Category	Ref.	Accuracy	Precision	Recall	F-measure	Scalability	Time	Cost
Lexicon-based	(Kauffmann et al., 2019)	✓		✓	✓			
	(Jiang et al., 2017)	✓						
Learning-based	(Dalla Valle and Kenett, 2018)	✓						
	(Jimenez-Marquez et al., 2019)	✓	✓	✓	✓			
	(Zhu et al., 2020)						✓	
	(Fan et al., 2020)	✓	✓	✓	✓			
	(Shirdastian et al., 2019)	✓	✓					
	(Lee and Paik, 2017)	✓				✓		
	(Sayed et al., 2020)	✓					✓	
Hybrid	(van Dieijen et al., 2020)	✓						
	(Spruce et al., 2020)	✓			✓			
	(Um et al., 2013)	✓	✓	✓		✓	✓	✓
	(Moise, 2016)					✓		
	(Hsu et al., 2017, 2017)	✓			✓	✓	✓	
	(Baltas et al., 2016)				✓			
	(Sun et al., 2018)	✓		✓	✓			
	(BalaAnand et al., 2019)	✓	✓	✓	✓	✓	✓	
	(Rodrigues and Chiplunkar, 2019)	✓	✓	✓	✓		✓	

which aimed at managing big data and processing UGC, built a Machine Learning Model (MLM). The second stage, which took MLM of stage one, involved a series of layers to build a big data architecture that analysed unstructured and heterogeneous data. The proposed framework was superior to its competitors in both quantitative and qualitative analysis. Despite high accuracy, better results may be obtained by applying the integration of advanced ML algorithms on different domains.

Despite the advancement and development of medical science, COVID-19 is the most perilous disease of the 21st century around the world, which is a critical threat to the physical and mental health of individuals. In this respect, Zhu et al. (2020) analysed the topics about COVID-19 in Weibo from January 24 to February 25, 2020. The authors tried to grasp the opinions of users about the epidemic from a temporal and spatial perspective in China. However, the study had some drawbacks. The spatial perspective of opinion analysis was limited to a provincial region. The age and gender of Weibo users were not considered, so they were not reflected in the analysis results. Moreover, since some users did not apply Sina Weibo to express their opinions, the result cannot be generalized. Thus, employing a high volume of data may lead to more predictive and accurate opinion analysis for relevant organizations in emergency conditions.

Fan et al. (2020) introduced a novel method for exploring real-time sentiments, team identification, and national identification of tweets during the 2018 FIFA world cup. The authors observed how the sentiments of fans' tweets in two matches (England vs. Croatia and England vs. Colombia) fluctuated during the match. They applied python and ensemble methods not only to design a model with high accuracy for sentiment analysis at different temporal points during the match, but also to analyse emojis as well as their valence. However, since 4% of the collected tweets were in Spanish and Croatian and the ML approaches cannot perform properly on Multilingual datasets, their method had low reliability. Moreover, they only analysed two English competitions, other international matches or other countries were not considered. Finally, all ML techniques did not have the ability to analyse the available sarcasm in tweets, so the results attained a low level of precision, recall, and F-measure.

Shirdastian et al. (2019) presented a framework to explore brand validity and their sentiment polarity both qualitatively and quantitatively. The authors explored opinion and sentiment polarity towards brand validity on Twitter dataset in terms of uniqueness, heritage, quality commitment, and symbolism. The study results indicate the enhancement of the proposed framework in precision and accuracy to find out the brand authenticity by exploring the related brand sentiments. The main drawback of this study was that neither was the variation of sentiments over time explored, and nor was the sentiment mining of bot-created brands excluded. Sayed et al. (2020) presented a hybrid approach that applied a combination of ML and lexicon techniques for sentiment analysis of tweets. The authors suggested a new metaheuristic approach based on Particle Swarm Optimization (PSO) and K-means to optimize data clustering. They evaluated their approach on four Twitter datasets with various topics employing spark streaming, leading to better accuracy in real-time analytics compared to previous approaches, but deep learning methods probably may lead to more accurate predictions.

To examine the relationship between volatility in the stock markets and UGCs, van Dieijen et al. (2020) presented a framework through the use of multivariate regression analysis and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model. The results showed the asymmetric impact of UGC on volatility, which means negative comments, compared to positive ones, increased volatility and had a significant effect on customers. For future research, scaling up may lead to practical implementation. Spruce et al. (2020) presented a new methodology for exploring the impact of social sensing and social data sentiment analysis of real-world events on named storms in the United Kingdom and Ireland. The authors collected tweets posted in winters 2017 and 2018. Then time zone, bot, and weather-related filters were applied to extract data related to weather incidents. By analyzing the sentiments of tweets during extreme climate events, the effects of weather incidents and their social impacts in terms of physical, emotional, spatial, and temporal perspectives were revealed and enhanced. The main limitation of this study was low scalability due to the small number of tweets retained in filtering weather-related tweets after the collecting phase. Further, the results were somewhat unreliable due to applying the python's sentiment analysis package (TextBlob) which has a training corpus based on movie review datasets.

Um et al. (2013) introduced a distributed and parallel parsing system based on MapReduce to analyse users' sentences in social sensor networks. To conduct the study, a Stanford parser with loose coupling was applied, which led to high scalability. Due to the parallel environment, the parsing time was low, the proposed system had high precision and high portability. The main limitation was that the actual data of social sensor networks like Twitter was not considered, and technical sentences were not analysed in the same way as ordinary users' phrases were. Moreover, researchers in (Baltas et al., 2016; Hsu et al., 2017; Lee and Paik, 2017; Moise, 2016) employed ML along with NLP for opinion and polarity mining of social big data in sentiment analysis that were applied for various decision-making purposes including marketing or health care issues like reporting drug side effects. In order to conduct sentiment analysis on a microblog big data platform, Sun et al. (2018) presented a model called Convolutional Neural Network-Long-Short Term Memory (CNN-LSTM). Each type of emotion was modeled through a Single Gaussian Model (SGM). The authors used CNN for extracting local attributes and LSTM as a global attribute extractor. The findings indicated that the sentiment of social language performed through CNN-LSTM model achieved high accuracy, but time was neglected in their model, and threshold selection was still taking too much time.

Also, BalaAnand et al. (2019) presented a mechanism to collect contents from social media by utilizing big sheets, big vision schemes, and sentiment assessment. In addition to Deep Learning Modified Neural Network (DMNN), which was used to investigate sentiments, the Modified Threshold-based Cuckoo Search Algorithm (MTCSA) was applied as a heuristic search algorithm for weight optimization. The experimental results revealed that the proposed Deep MNN outperformed in terms of reliability, robustness, scalability, accuracy, precision, recall, F-measure, and computational time in comparison with other algorithms, but the cost of the proposed method was not assessed. For topic classification and sentiment analysis of social big data, Rodrigues and Chiplunkar (2019) presented a distributed Hadoop framework. Additionally, the Bag-of-words method was used to classify the relevant tweets into six different groups. Then four various NLP methods, namely Lexicon uni-gram, bi-gram Lexicon, uni-gram NB, bi-gram NB, and Hybrid Lexicon-Naive Bayesian Classifier (HL-NBC), were employed. HL-NBC was more effective and outperformed other classifiers in terms of accuracy, execution, and response time. However, separating and classifying sarcastic sentences and cross-lingual opinions for sentiment analysis were still unsolved challenges.

4.2. Network-oriented approaches

Network-oriented approaches analyse big social data based on nodes or entities and their relations within social networks. Network-oriented approaches are classified into two groups: Embedding learning and community learning. We review the selected papers with embedding learning and community learning approaches in Sections 4.2.1 and 4.2.2, respectively. In Sections 4.2.1 and 4.2.2, the classification of techniques, the definition of methods, and the related papers are discussed.

4.2.1. Overview of the embedding learning approaches

Some of the reviewed papers presented embedding learning that focused on extracting valuable information about users and nodes inside a network for link prediction, influence analysis, and information diffusion in social networks. Social influence means an individual's ability to influence another user in a network; the more influential a person is, the more followers he will have (Kumaran and Chitrakala, 2017). The embedding learning approach aims to analyse a network based on users and their features and model the process of information diffusion on online social networks through learning user's characteristics and dissemination of information among users. Embedding learning approaches try to find the influence of different nodes in a network by identifying the position of a node in a path or a number of paths in which it occurs; the node that is most often in the center of a network and has more paths is more influential.

In the aspect of predicting the underlying diffusion process, three categories are distinguished in embedding learning approaches: Graph-based, non-graph based, and explanatory. Graph-based and non-graph based are kinds of predictive models in which, by investigating the previous information propagation, the information dissemination is predicted from spatial or/and temporal points of view. Graph-based approaches focus on the static and graphical structure of a network in which information is transmitted and predicts who influences whom. In this approach, each node can be activated or deactivated, such as Independent Cascades (IC) and Linear Threshold (LT), while in non-graph based approaches, the topology and structure of a network are not taken into account and each node is randomly connected to other nodes in the network with an equal probability such as epidemic models, Linear Influence Model (LIM) and Partial Differential Equations (PDEs). The main goal of explanatory models is to infer the information propagation path and to show how the information is propagated in social networks. Propagation characteristics such as pairwise transmission rate, pairwise transmission probability, and cascade properties are explored in this model whereas the network in which information diffusion takes place is unknown.

This section presents the selected papers with embedding learning approaches. In addition, the selected papers that use this approach in social big data analysis are reviewed. Finally, they are compared and summarized in Tables 7 and 8. Table 7 compares them in terms of main ideas, advantages, disadvantages, evaluation methods, tools, and case studies along with their categories. In some studies, the applied tools for analyzing and implementing the intended approach were not mentioned. The evaluation parameters are also specified in Table 8.

Kumaran and Chitrakala (2017) offered a social influence method based on rank-sampling approach. After collecting Twitter's data, parallel information diffusion modelling, which took the users' queries as input, determined forwarding nodes and calculated the path of information flow. The next portion was influential spreader ranking, which took a search query and applied topological and users' attributes to calculate users' feature scores. At last, two solutions were provided for an influence maximization problem.

Table 7
Reviewing and comparing papers with embedding learning approaches.

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
Graph-based	(Kumaran and Chitrakala, 2017)	Introducing a social influence rank-based determination method on big data streams in online social networks	• High scalability • High accuracy over time • Decreasing running and computation time	• The fixed size of the sample • Additional features could be added in future	Data sets	Python,Hadoop, MongoDB	Twitter
	(Persico et al., 2018)	Introducing an influence maximization and diffusion algorithm	• High scalability • Reduce running time • Low cost • High accuracy • High recall • Supporting SN applications	• Not evaluating the performance of these two architectures on other datasets • Low privacy-preserving	Real test bed	Apache Storm, Apache Spark, Microsoft Azure HDInsight	Yahoo Flickr Creative Commons 100 Million (YFCC100M)
	(Gao et al., 2017)	Presenting an information-dependent embedding based diffusion prediction model	• High precision • Efficient diffusion • Prediction speed (low response time)	• Not considering the social structure in the proposed embedding model	Real test bed	Not mentioned	Digg,Meme tracker, GOOGLE +
	(Elkin et al., 2017)	Introducing a network-based model to predict disease activity across geographical locations	• Being able to predict disease and helping to control diseases • High accuracy • High F-measure	• Low scalability • Not considering other factors besides geographic locations such as weather patterns	Real test bed	Not mentioned	Twitter
	(Wang et al., 2014)	Presenting a heuristic approach to maximize influence in social networks	• Low running time • High scalability	• Low privacy-preserving • Not measuring the accuracy of the model	Simulation	Not mentioned	Political blogs, Netscience dataset
	(Talukder and Hong, 2019)	Proposing a heuristic model for minimizing viral marketing costs in social networks	• Low cost • Low running time	• Not analyzing the complexity order or the price of the proposed model	Simulation	Python	Facebook, Epinions
	(Chen et al., 2020)	Presenting a topic-aware influence maximization model based on cloud computing	• High computational efficiency • Low running time • Requiring low storage compared to previous methods (Low cost) • High scalability	• Not evaluating the accuracy of the proposed model by various evaluation parameters	Simulation	Not mentioned	NetHEPT,Epinions,DBLP, LiveJournal,Friendster
Non-graph based	(Wu et al., 2020)	Offering a protection and recovery model, examining the influential users, and studying virus propagation	• Low running time • Low cost	• Fixed number of nodes and connections	Simulation	Matlab	Undirected network BlogCatalog and directed network As-level network
	(Wu et al., 2018)	Proposing an algorithm and calculation model for searching the relationship between nodes, big data, and small data	• Low cost (complexity reduction) • Improve the delivery ratio (response time)	• Time (Information delivery time)	Simulation	Not mentioned	Population map of Beijing city in China
	(Wu et al., 2018)	Presenting mobile nodes to explore and limit the spread of rumors in social networks	• Detecting and preventing the spread of rumors	• Not able to reduce the spread time point of the rumors earlier	Simulation	C#	Facebook
			• High security		Simulation	C#Simulator	

(continued on next page)

Table 7 (*continued*)

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
	(Peng et al., 2017)	Introducing an immunization framework for mobile social networks		• Unable to detect social network malware in real-time • High time complexity			Largest Cellular Network in China
Explanatory	(Óskarsdóttir et al., 2019)	Proposing a financial credit scoring model which uses mobile phone data and social network analytics	• High accuracy	• Low privacy-preserving • Analyzing data only for one type of credits and a single country	Data sets	Not mentioned	CDR data of cell phone numbers and the data bank of customers that both operate in the same country
	(Raj and Babu, 2015)	Proposing mathematical models to compute the probability of staying in social network and FIAEC	• Low cost • Increasing the number of connections and interactions between the connections	• Low scalability	Real test bed	Not mentioned	Facebook
	(Su et al., 2016)	Introducing a framework to deliver mobile social data over content-centric mobile social networks	• Low delay	• Low scalability • Not considering dynamic mobile social users • Low privacy and security protection • Limited resource allocation such as bandwidth, and buffer space	Simulation	Not mentioned	Not mentioned
	(Kumar et al., 2016)	Recognizing the influential user on Twitter by applying the number of followers and friends	• High scalability	• High execution time	Real test bed	R, Hadoop, Python	Twitter
	(Zhang et al., 2017)	Analyzing real-world device-to-device datasets in mobile social networks	• High scalability • High speed • Parallel processing of data	• Low security and low content privacy	Real test bed	Apache Spark, Apache Kafka, Hadoop	Not mentioned
	(Xu et al., 2015)	Analyzing the impact of various sampling approach on the influence diffusion on social big data	• Low cost • High accuracy	• Low scalability	Real test bed	Not mentioned	Twitter
	(Yang et al., 2020)	Suggesting a depression detection framework by applying ML techniques	• High scalability • High privacy-preserving	• Not considering the topics that users posted • Not considering various genders, age groups, and their depression risk level	Data sets	Apache Spark, R programming language	Facebook
	(Maireder et al., 2017)	Proposing two social network measures of communicative activities to characterize information diffusion	• Being able to understand the information diffusion of political contents within a social network by two proposed measures • Advancement in communication patterns within online social networks	• Not analyzing the contents and types of tweets and messages throughout the network • Not developing on different topics and contents	Real test bed	Gephi (Network analysis software)	Twitter discussion of TTIP in Europe

Table 8
An overview of the evaluation parameters in papers with embedding learning approaches.

Category	Ref.	Accuracy	Precision	Recall	F-measure	Scalability	Time	Cost	Influence Diffusion	ROC (AUC)	Kappa	Security
Graph-based	(Kumaran and Chitrakala, 2017)	✓				✓	✓		✓			
	(Persico et al., 2018)	✓		✓		✓	✓	✓				
	(Gao et al., 2017)		✓				✓		✓			
	(Elkin et al., 2017)	✓			✓							
	(Wang et al., 2014)					✓						
	(Talukder and Hong, 2019)						✓	✓				
	(Chen et al., 2020)					✓	✓	✓				
Non-graph based	(Wu et al., 2020)						✓	✓				
	(Wu et al., 2018)						✓	✓				
	(Wu et al., 2018)						✓					
	(Peng et al., 2017)								✓			✓
Explanatory	(Óskarsdóttir et al., 2019)	✓										
	(Raj and Babu, 2015)							✓				
	(Su et al., 2016)						✓					
	(Kumar et al., 2016)					✓						
	(Zhang et al., 2017)					✓	✓					
	(Xu et al., 2015)	✓						✓	✓			
	(Yang et al., 2020)	✓	✓	✓	✓					✓	✓	
	(Maireder et al., 2017)								✓			

Ranking-based sampling, MapReduce, and parallel processing were applied to ensure accuracy and time reduction, respectively. Despite scalability, the sample size was considered fixed, so an approach that could define the most appropriate sample size was needed to be performed.

In another research, Persico et al. (2018) analysed the efficiency of two big data architectures, namely Lambda and Kappa. Although the size of the dataset affects the performance, both architectures provided good scalability, but in case of increasing input size, Lambda had higher performance than Kappa due to its in-memory computation. Findings indicated that the deployment for Kappa with the same number of executors was more expensive than Lambda. Besides, in both architectures, the performance was improved when the algorithm was executed on more massive clusters. In case of virtual machines (VMs) characteristic enhancement (or with resource-richer nodes), Kappa significantly improved the performance (vertical scaling). In general, reports showed that Lambda performed better, and both architectures supported social network applications properly. To predict information diffusion in the content of social big data, Gao et al. (2017) offered an efficient Information-dependent Embedding Based Diffusion Prediction (IEDP) model. They also extended a typical margin-based optimization algorithm and presented an efficient learning algorithm based on Stochastic Gradient Descent (SGD). The complexity of the proposed model was significantly reduced, but the social structure was not considered in their proposed embedding model.

Additionally, for illness control and prediction in advance, Elkin et al. (2017) introduced a network-based approach for modeling illness activity and generated predictions about ILI based (Influenza-Like Illnesses) across geographical locations. This prediction model could help with illness control and provided predictions for one week in advance. Meanwhile, it was unsuccessful with airline traffic data in predicting ILI activity across geographies and had a low level of scalability, and except for geographical locations, other factors such as weather patterns or low population density were not considered. By discovering more factors, the model could have been stronger. Moreover, a heuristic model called PRDiscount was proposed in (Wang et al., 2014) to select the first seeds for maximizing the influence diffusion in social networks. On the contrary, Talukder and Hong (2019) introduced a heuristic mixed approach to minimize and optimize viral marketing costs in social media.

Since nowadays social networks have a great impact on the dissemination of information and users' comments and on individuals' daily lives, Chen et al. (2020) suggested a topic-aware influence maximization model based on cloud computing. They employed a sketching technique along with a greedy algorithm to discover the optimal top-k seed users that maximize the influence of information being spread within a network. Compared with available influence maximization approaches, the proposed approach achieved low running time and low storage, but a limited number of evaluation parameters were applied to verify the accuracy of the model.

Moreover, to discover the influential users, Wu et al. (2020) offered a Protection and Recovery Strategy model (PRS) to study the propagation of the virus in social networks. In the proposed mechanism, the users were divided into five groups based on their reactions to the virus: Susceptible, Contagious, Doubt, Immune, and Recoverable (SCDIR). The PRS model made it possible to control viruses and to reduce infected users. Despite the low running time and low cost of the model, a fixed number of nodes and connections were assumed; the dynamic changes in a number of nodes and their connections may lead to different results. Wu et al. (2018) suggested a model to search small data and to compute the effect of small data nodes to use them instead of big data. They believed that obtaining small data leads to a reduction in the complexity of big data. Results showed that 1% of small data could connect 15% of communication nodes, and 20% of small data could broadcast 80% of data packets, so the other nodes were in waiting status. Although complexity was decreased and the delivery ratio was improved, a new algorithm was needed to establish a trade-off between reliability, delivery ratio, delay, and the use of limited network resources.

Wu et al. (2018) presented a developed model to recognize and restrict the process of rumor dissemination among users by considering all the users' behaviors. A time threshold was dedicated to each user to indicate the delays in users' reactions. The authors suggested a mobile node to propagate authorized information to decrease the penetration of rumors. They simulated the proposed model on the Facebook dataset to investigate the influence of speed, arrival time, and strategies of the mobile node on rumors. The speed and the strategy of mobile nodes could not reduce the spread time point of rumors earlier, but in general, it reduced the spread time of rumor; therefore, the best solution to detect rumors is to send mobile nodes to neighbor nodes with the highest degree.

Furthermore, to prevent the spread of malwares, Peng et al. (2017) presented a big data-based framework in which social interactions were transformed into a bidirectional weighted graph that displayed people's daily SMSs/MMSs. Moreover, social influence, involving direct and indirect influence, was measured. Then a set of immunization algorithms were designed, and the Susceptible Infectious Recovery (SIR) model was developed because the top k influential nodes had more influence on the distribution of malware propagation. Thus, based on the presented immunization strategy, the top k influential nodes were minimized; meanwhile, it did not detect social media malware in real-time.

In order to improve the statistical and economic performance of credit scoring applications both, Óskarsdóttir et al. (2019) employed personalized Page Rank (PR) and SPreading Activation (SPA) methods on Call-Detail Records (CDR), credit and debit account information. The results showed that the features of calling behavior were most effective, and the information extracted from CDR data in terms of "value" facilitated financial prediction. The major challenge was how to maintain privacy-preserving of customer's data. Moreover, only one type of credit was analysed; other types of credits may lead to different results.

Furthermore, Raj and Babu (2015) proposed Firefly Inspired Algorithm for Establishing Connections (FIAEC) and mathematical models for computing the probability of staying in social networks. The goal of this algorithm was to maximize the number of connections concerning *n* individual in social network sites. By using the proposed algorithm, the number of connections was increased, and so did the interaction between connections. On the other hand, FIAEC was not scalable, and it was only tested for a sample size of 10,200 and 600.

Su et al. (2016) studied the characteristics of mobile big data and presented a new framework to spread these data over content-centric Mobile Social Networks (MSNs). To resolve volume, variety, control, and manage mobile big data challenges, the framework was delivered over CCNs. Findings showed that a low value of weight coefficient for a data packet led to a low delay. As their proposed framework was based on static characteristics, it did not consider dynamic mobile social users and was tested on a limited number of users, so it was not scalable. The limited resource allocation, such as bandwidth and buffer space, was not considered, and security was not maintained for the data stored out of their own mobile devices. In addition, to recognize the influential users, Kumar et al. (2016) developed a methodology by applying the number of friends and followers of accounts. In another study, Zhang et al. (2017) analysed an offline device-to-device dataset in mobile social big data and pushed interesting contents to the most influential users.

Besides, Xu et al. (2015) investigated the impact of various sampling approaches on the distribution of tweets and measured retweets to identify the influence diffusion in social network analysis. Since a notable amount of data in social networks are related to people who declare their opinions and thoughts, Yang et al. (2020) offered a social big data analysis framework to diagnose depression efficiently. The authors applied a large Facebook dataset to evaluate the proposed framework by investigating the effect of both friendship influence and users' intentions and interactions on users' mental health. They evaluated the performance of the framework with a various subset of social and user-level features to indicate that the users' social interactions with their friends on social networks could show their mental states. Unlike other researchers, to analyse friendships' influence, both indirect and direct neighbors of a user were investigated; however, the topics of users' posts were not considered as well as various genders, age groups, and their depression risk level.

Additionally, in order to investigate the diffusion structure of networks, Maireder et al. (2017) presented two new social network measures, namely Audience Diversity Score (ADS) and Communication Connector Bridging Score (CCBS). ADS identified the diversity of a particular actor's followers, and CCBS highlighted the account that bridge and diffuse information throughout the entire network. The results demonstrated that the network was not divided by a unique factor but by a set of influential ones, like language, geo-identity, and political trends. Despite the advancement in communication patterns, the contents and types of tweets broadcast across the network were not analysed. Moreover, ADS and CCBD measures were not combined to detect the two-factor interaction in the spread of information.

4.2.2. Overview of the community learning approaches

As we stated earlier, social networks comprise a set of vertices or nodes in which nodes stand for users and individuals, which are associated with one another through numerous edges that represent their relations and interactions (Leung and Zhang, 2016). "Community" is referred to as groups of individuals who have similar interests, attitudes, or common characteristics (Wu et al., 2018). From the social aspect, detecting groups of individuals in a network on structural and topological properties is known as community learning which is crucial for various perspectives in society such as business and recommendation systems. Thus, it leads to innovative approaches for identification of communities that can be carried out in micro (micro-communities) or macro (macro-communities) network structural features. In community detection, the assumption is that people in one community interact more with one another because of the similarity of interests among them compared with other communities, so the network is divided into various communities.

In community learning, after identifying clusters of nodes, the number of clusters is determined. A cluster is mapped into a community, then the probability distribution over interactions among users and also within and among clusters is estimated. Community learning approaches can be categorized into node-based or group-based approaches to recognize the communities. Node-based approaches are carried out based on the properties of network nodes. Since similar nodes belong to the same communities, node degree, node similarity, or node reachability are considered in this approach. While group-based approaches do not regard characteristics at the node-level and consider the characteristics and the connections of the whole group and network by recognizing balanced, robust, modular, dense, or hierarchical communities.

In this section, the selected papers with community learning approaches are reviewed. Table 9 depicts a comparison of the selected papers with community learning approaches. It includes the main ideas, advantages, disadvantages, evaluation methods, tools, and case studies along with their categories. Table 10 shows the parameters that these papers with community learning approaches have used to evaluate their methods.

Aksu et al. (2013) presented a multi K-core and multi-resolution solution for social network community detection. The authors offered a distributed and scalable algorithm that ran on Apache HBase to compute K-core subgraphs for both client and server-side. The experimental results on dynamic networks indicated that despite such advantages as robustness, parallel, and distributed processing, the proposed algorithm was very costly in case of inserting and deleting edges. Wu, et al. (Wu et al., 2018) presented a hash-based approach along with graph mining to discover interactions and communities among users in social media in which a trade-off between efficiency and effectiveness of incremental and time slices-based approaches was guaranteed.

Since the result of SNA helps managers in decision making for their markets, Dabas (2017) considered an electronic store with 98 employees, who were responsible for selling, maintaining, and installing mobile phones, tablets, and so on. For experiments, Pajek and different metrics of SNA like degree centrality, betweenness centrality, stress centrality, Power Centrality (PC), Information Centrality (IC), reachability matrix, and clustering coefficient were used. The social analysis informed executive managers of customers' reactions in real-time to respond quickly if necessary, but it suffered from inadequate security of sensitive and personal data. While Yousfi et al. (2016) proposed a solution to construct the graph of social big data to enhance the semantic extraction by graph analysis.

As finding the right researcher with the best experience and knowledge is time-consuming and critical in research communities, Sun et al. (2015) presented an expert recommendation method based on topic relevance, expert quality, and researcher connectivity for experts in scientific communities. The architecture of this expert finder system contained three phases (profiling, modeling, and ranking). Large-scale computation task was supported as well as linear speed up and high accuracy. In their method, except for AHP in the ranking phase, the authors did not use other techniques as the rank aggregation model. In another study, to enhance the quality of vehicle localization in vehicular networks, Lin et al. (2016) proposed an Overlapping and Hierarchical Social Clustering (OHSC) model. The OHSC model explored the social relations between vehicles, and then classified the vehicles into different social clusters. As

Table 9
Reviewing and comparing papers with community learning approaches.

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
Node-based	(Aksu et al., 2013)	Presenting a multi-resolution community detection algorithm on the Hadoop platform	• Low response time • High scalability • Robustness • Parallel processing	• High cost	Real testbed	Java, Hadoop, Apache HBase	Orkut,LiveJournal,Flickr, Patents,Skitter,BerkStan, YouTube,WikiTalk,Dblp
	(Wu et al., 2018)	Proposing an incremental community detection model	• Low time • Low cost	• Not being tested on other algorithms except for K-clique • Low scalability	Real testbed	Not mentioned	DBLP (Digital Bibliography & Library Project Dataset)
	(Dabas, 2017)	Presenting big data analytics for exploratory social network analysis	• Low response time	• Low security • Low privacy-preserving	Real test bed	Pajek	An electronic store with 98 employees
	(Yousfi et al., 2016)	Presenting a cloud-based service to manage social big data	• High scalability	• High cost	Prototype	MySQL, Apache Hadoop, GraphLab(Java), Apache Flume	Twitter
Group-based	(Sun et al., 2015)	Introducing an expert finder system based on big data analytics	• Linear speed up (Response time) • High accuracy • High scalability	• Not considering other data fusion techniques, and using only one method for modeling score distribution • Considering a small number of human factors	Prototype	Hadoop	Scholar Mate
	(Lin et al., 2016)	Proposing a social based localization algorithm and OHSC model	• High accuracy • High security	• Low stability • Low reliability	Simulation	Java SE development	Not mentioned
	(Kuang et al., 2016)	Introducing a tweet ranking model	• High precision • Proposing a ranking model that improves tweet ranking performance	• Considering a small number of indicators for ranking by analyzing users' behaviors	Real test bed	Not mentioned	Sina microblog
	(Jin et al., 2015)	Designing a distributed community structure mining framework by using MapReduce	• High accuracy • High precision • High recall • High scalability • Low cost	• High running time	Real test bed	Hadoop	Large-scale artificial dataset,Real-world social media networks
	(Li et al., 2016)	Presenting a cloud-based online learning algorithm for social big data analysis	• High privacy • High scalability • High accuracy • High regret bound	• High delay	Simulation	Hadoop	Not mentioned
	(Paik et al., 2017)	Proposing a parallel approach for creating a graph network	• Low execution time • High scalability • High precision • High recall	• High cost	Real test bed	Hadoop	Not mentioned
	(Karimi et al., 2018)	Analyzing defrauding information in social networks by employing Apache Hadoop	• Real-time execution • High scalability	• High cost	Real test bed	Apache Hadoop, Gephi, Apache Nifi, Apache Solr	Twitter

(*continued on next page*)

Table 9 (*continued*)

Category	Ref.	Main ideas	Advantages	Disadvantages	Evaluation methods	Tools	Case studies
	(Leung and Zhang, 2016)	Proposing a method to represent and manage social big data	• High scalability • Low running time • Low space requirement (low cost)	• Low privacy-preserving	Real test bed	Apache Hadoop, Java	The Stanford Network Analysis Project(SNAP) ego-Facebook, ego-Twitter dataset
	(Sharma, 2018)	Presenting a real-time framework for analyzing Twitter data by applying graph analysis	• High reliability	• Not considering the edge properties in algorithms • Not evaluating the tolerance value	Real test bed	Apache Spark	Twitter,Sina Weibo, Tencent Weibo
	(Du, 2018)	Offering SNA method by adding semantics into nodes and edges in the weighted undirected graph	• High accuracy	• Not applying additional information to nodes or edges to improve graph analysis	Real test bed	Not mentioned	Dow Jones Industrial Average (DJIA), Stock exchange markets (NYSE and NASDAQ)
	(Wang et al., 2017)	Presenting a U-model for directed and undirected graph based on similarities	• The high clustering coefficient of the proposed model • Low cost • High scalability	• Not analyzing the statistical characteristics • Not investigating the impact of the proposed model on the information diffusion process	Simulation	Not mentioned	Sina Weibo, Tencent Weibo, Twitter
	(Ghosh et al., 2016)	Offering a fuzzy logic and density-based clustering algorithm for big data analysis	• High scalability	• High complexity	Real test bed	Not mentioned	Facebook,YouTube
	(Wang et al., 2017)	Analyzing entrepreneurial social big data	• Detecting community	• Not analyzing entrepreneurial networks in more sparsely populated areas	Real test bed	MongoDB	Twitter

Table 10
An overview of the evaluation parameters in papers with community learning approaches.

Category	Ref.	Accuracy	Precision	Recall	Scalability	Time	Security	NMI	Cost	Centrality Measures	Clustering Coefficient
Node-based	(Aksu et al., 2013)					✓			✓		
	(Wu et al., 2018)	✓				✓			✓		✓
	(Dabas, 2017)					✓				✓	✓
	(Yousfi et al., 2016)				✓						
Group-based	(Sun et al., 2015)	✓			✓	✓					
	(Lin et al., 2016)	✓					✓	✓			
	(Kuang et al., 2016)		✓			✓					
	(Jin et al., 2015)	✓	✓	✓	✓			✓	✓		
	(Li et al., 2016)	✓			✓		✓				
	(Paik et al., 2017)		✓	✓	✓	✓					
	(Karimi et al., 2018)				✓					✓	
	(Leung and Zhang, 2016)				✓	✓				✓	
	(Sharma, 2018)					✓					
	(Du, 2018)									✓	✓
	(Wang et al., 2017)				✓					✓	✓
	(Ghosh et al., 2016)				✓						
	(Wang et al., 2017)									✓	

a result of OHSC, a Social based Localization Algorithm (SBL) was presented to support the global localization through vehicle location prediction even without the GPS devices. Although SBL had a high overall performance in the vehicle localization, the SBL algorithm had low stability and the worst performance in location error.

By increasing active users and daily tweets, users are faced with a severe problem of overloading information. To overcome ranking and recommending challenge, most micro-blogging services organize tweets in a timely order that place newer tweets at the top, but all these tweets may not be attractive to users. Kuang et al. (2016) proposed a new tweet ranking model considered three main aspects, consisting of the popularity of a tweet itself, the intimacy between the user and the tweet publisher, and the user's interest areas. This ranking model improved tweet ranking performance; however, more indicators for ranking in analysing users' behaviors were not considered. In order to identify all hidden communities in social media networks, Jin et al. (2015) designed a framework for community structure mining in which network partitioning process was avoided, and map equation process ran directly on MapReduce in the new framework. Instead of PageRank, the authors employed local information of nodes and their neighbors for calculating the distribution probability related to each node. The framework outperformed the previous algorithms, such as Radetal and FastGN, in accuracy, velocity, and scalability. However, the greedy search method that was applied to find an appropriate node for combining had some limitations that needed to be improved.

Additionally, Li et al. (2016) offered a distributed algorithm for data centers to handle social data to ensure privacy and guarantee the prediction accuracy improvement in real-time. Further, Paik et al. (2017) presented an effective service discovery through the creation of a graph-based algorithm based on MapReduce and parallel programming. In (Karimi et al., 2018), Twitter data were analysed, and the degree centrality was calculated to investigate deceiving information based on a parallel approach. Leung and Zhang (2016) offered a novel method to represent and manage social big data. They employed graph mining approaches in directed, bi-directed, undirected, and bipartite graphs for analyzing and mining social big data in distributed settings. In (Sharma, 2018), researchers designed a framework to analyse real-time Twitter hashtags by employing hashtag co-occurrence graph and connected components algorithm. Moreover, Du (2018) developed a high-frequency pair trading algorithm to perform semantic analysis on a weighted undirected graph by employing SNA approaches along calculating centrality parameters in a stock market.

Since similar nodes are usually placed in the same cluster, in (Wang et al., 2017), a U-model was introduced for directed and undirected graphs based on similarity, which could define social big data characteristics, clustering coefficient, degree, and distance distribution accurately. In order to analyse the conversation in a social network, Ghosh et al. (2016) offered a new algorithm utilizing fuzzy methodology and density-based clustering on social clouds. This study was applied to examine the rate of users' participations to find the popularity of the subject under discussion. Besides, this algorithm could have been developed towards more heuristic-based graph mining and put a benchmark towards heuristic optimization. Further, to represent the structures of network communities, Wang et al. (2017) digitally analysed Twitter's data about diverse actors involved in entrepreneurial networks by applying the Clauset-Newman-Moore algorithm. The counties that were in the same cluster had stronger internal interactions than those in different clusters, but this research did not analyse entrepreneurial networks on Twitter data and in case of lacking the participation of users in low population regions of the country.

5. Analysis of results

The results of this systematic review are analysed in this section. Section 5.1 presents an overview of the selected papers. Since the goal of this review is to highlight the differences, advantages, and disadvantages of various big data analytic approaches in social networks, a discussion of the mentioned classification is outlined in Section 5.2.

5.1. Overview of the selected studies

The following complementary questions are defined to explore the state-of-the-art on big data analytic approaches applied in social networks.

- Which publishers have published most papers on big data analytic approaches applied in social networks?
- How was the distribution of publishers and studies per year on big data analytic approaches applied in social networks?
- How was the distribution of studies per publication channel on big data analytic approaches applied in social networks?

In this section, the distribution of 74 papers reviewed in Section 4—categorized by publishers, the year of publication, the number of papers by year, and the percentage of papers classified by publishers—is shown in Figs. 5–7, respectively. Fig. 5, which states the papers over time, indicates that ScienceDirect, and Inderscience, have published papers in this field since 2015. IEEE, Springer, and ScienceDirect have provided the highest number of papers in this area, respectively. Also, Emerald and Taylor&Francis have presented

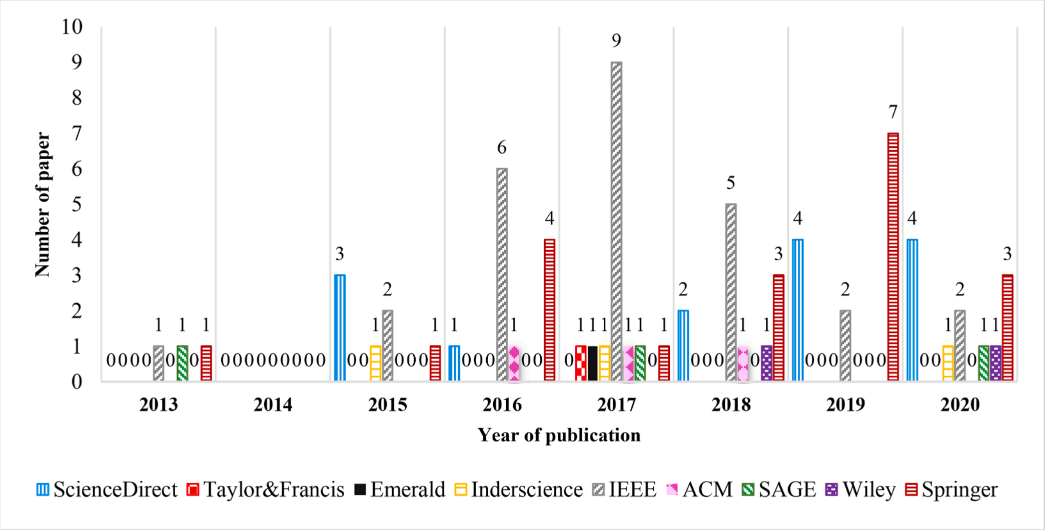

Fig. 5. The number of the studied papers categorized by publishers and years.

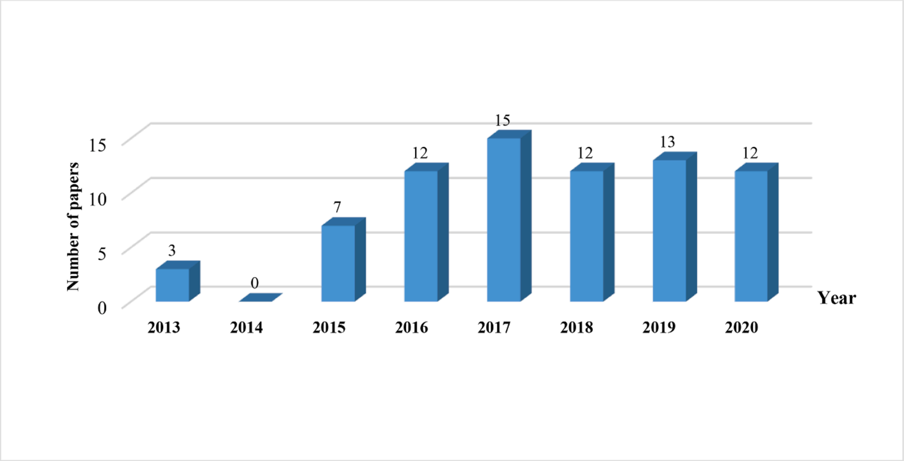

Fig. 6. The number of the studied papers by years.

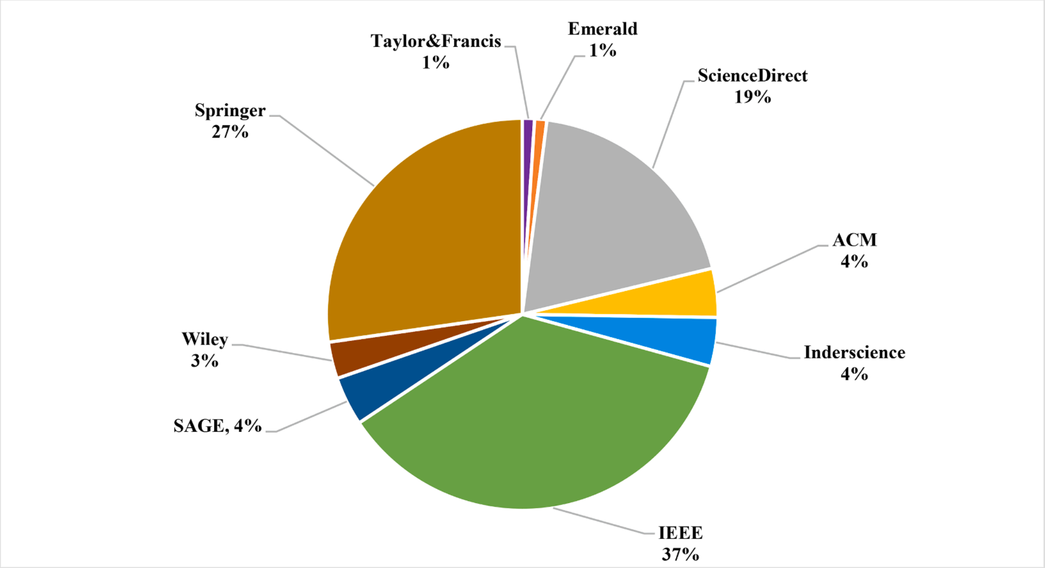

Fig. 7. Percentage of the studied papers categorized by the publishers.

Table 11
Distribution of the studies per publication channel.

Category	Publisher	Publication channel	Count
Journals	**IEEE**	IEEE Access	4
		IEEE Transactions on Multimedia (TMM)	2
	Science direct	International Journal of Information Management (IJIM)	3
		Industrial Marketing Management (IMMGT)	2
		Future Generation Computer Systems (FGCS)	2
	Springer	Multimedia Tools and Applications (MTAP)	3
		Wireless Personal Communications (WPC)	2
Conferences	**IEEE**	IEEE International Conference on Big Data (Big Data Congress) (IEEE Big Data)	3
		International Conference on Circuits, Controls, Communications and Computing (I4C)	2

the least number of papers. Fig. 6 shows that most papers in this subject were published in 2017 and 2019. Fig. 7 illustrates the classification of papers among nine publishers, out of which IEEE and Springer have provided 37% and 27% of the papers, respectively. 19% of the total papers were related to ScienceDirect, while, ACM, Inderscience, and SAGE publishers had 4% of the papers each. Also, 3% of the papers were published by Wiley. Additionally, Taylor&Francis, and Emerald, had 1% of the reviewed papers each.

In Table 11, we demonstrate the distribution of publication channel that published more than one paper among 74 studied papers. Table 11 depicts that 23 papers were published in IEEE Access (IF = 3.745), TMM (IF = 5.452), IJIM (IF = 8.210), IMMGT (IF = 4.695), FGCS (IF = 6.125), MTAP (IF = 2.313), WPC (IF = 1.061), I4C, and IEEE Big Data.

5.2. Research objectives, approaches, and evaluation parameters

The reviewed studies were studied and classified according to various characteristics to answer some of the research questions listed in Section 3.1, as explained below:

- Q1: What are the existing big data analytic approaches applied in social networks?

Big data analysis has many applications in social networks and is performed in various ways. As it was stated earlier, selected papers were reviewed, and big data analytic approaches in social networks were described in two main categories based on their analysis method: Content-oriented approaches, and network-oriented approaches. In content-oriented approaches, user-generated posts are analysed with the aid of lexical codes, linguistic codes, and statistical tools. Meanwhile, network-oriented approaches considered nodes or users and their relations for big social analysis. Also, the interaction between social group members and the relationship between group members and people outside the group are discovered. We categorized content-oriented approaches into two groups, topical learning and opinion/sentiment learning, and network-oriented approaches into two groups: Embedding learning and

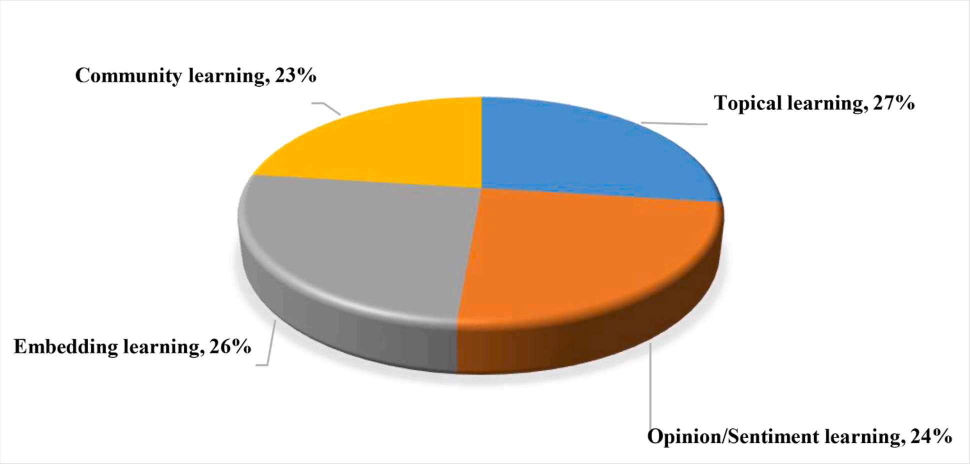

Fig. 8. Percentage of social big data analytic techniques in the selected papers.

Table 12
A summarization of the advantages and disadvantages of the discussed taxonomy.

Approach	Category	Advantages	Disadvantages
Content-oriented approaches	*Topical learning*	• Better accuracy • Better precision • Better recall • Decreasing cost • Decreasing execution and response time	• Unacceptable privacy-preserving • Unacceptable security
	Opinion/Sentiment learning	• Better accuracy • Better precision • Better recall • Better F-measure • Decreasing execution and response time	• Unacceptable cost
Network-oriented approaches	*Embedding learning*	• Decreasing execution and response time • Decreasing cost • Better accuracy • Better scalability	• Unacceptable privacy-preserving
	Community learning	• Better accuracy • Better scalability	• Unacceptable security • Unacceptable privacy-preserving • Unacceptable cost

community learning.

Fig. 8 represents the percentage of social big data analytic techniques in reviewed papers based on Fig. 4. Fig. 8 shows that the content-oriented approaches have the highest percentage (51%) in which topical learning and opinion/sentiment learning comprise 27% and 24% of the studied papers in the literature, respectively. Further, 49% of the papers are network-oriented approaches out of which 26% and 23% of the papers are related to embedding learning and community learning, respectively. The main properties of the selected papers reviewed were shown in Tables 3, 5, 7, and 9. The selected papers were evaluated based on critical parameters such as accuracy, scalability, precision, recall, F-measure, cost, and time. The advantages and disadvantages of the discussed taxonomy are summarized in Table 12 based on Tables 3, 5, 7, and 9. As specified in Table 12, the main focus of researchers in content-oriented approaches are on some parameters such as accuracy, precision, recall, and time. This table also illustrates that accuracy and scalability are enhanced in network-oriented approaches, but privacy and security are not considered by most researchers. Moreover, findings have shown that since manipulating community-based features is challenging and not user-controlled, and extracting these features requires an in-depth analysis of a large and complex social community, which has high complexity and requires plenty of resources, community learning approaches have high costs. Besides, according to Table 12, security and privacy-preserving are still the main drawbacks of community learning approaches.

- Q2: What parameters do the researchers employ to evaluate the big data analytics in social networks?

In this study, reviewed papers have been evaluated by various evaluation parameters, which were presented in Tables 4, 6, 8, and 10. Fig. 9, illustrates the parameters used by researchers to evaluate the techniques and methods applied in reviewed papers. The

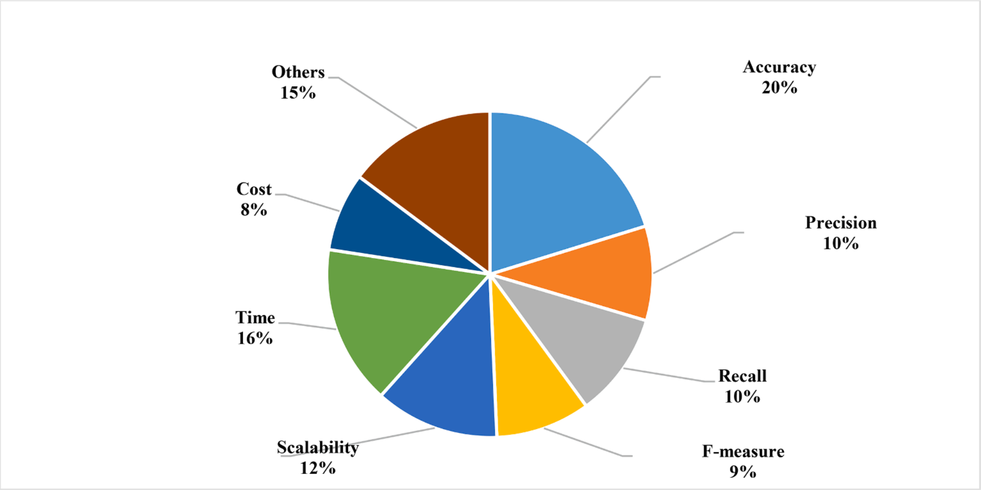

Fig. 9. Percentage of evaluation parameters in the selected papers.

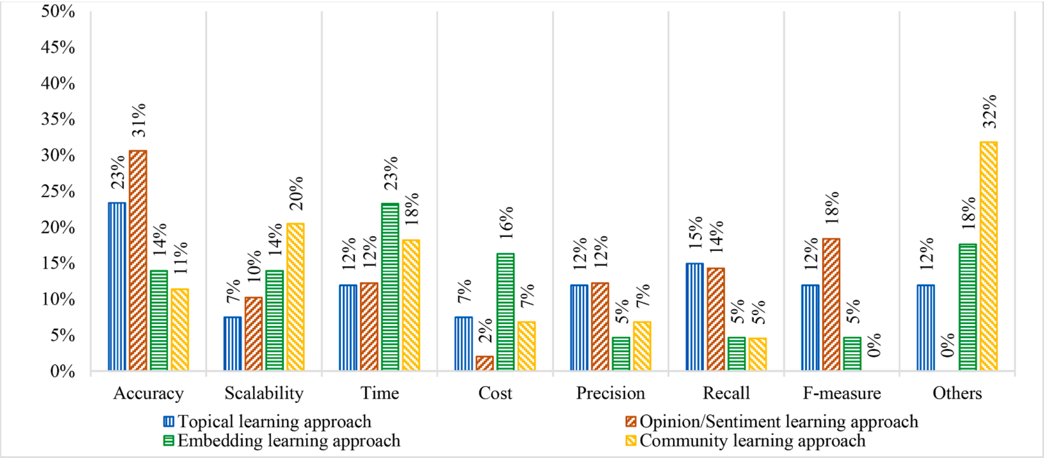

Fig. 10. Percentage of evaluation parameters in each approach of the selected papers.

results of the provided comparison in Fig. 9 show that 20% of the studies have enhanced accuracy, 16% of them have reduced time, and 12% of the studies have assessed scalability. Recall, precision, F-measure, and cost were also important among parameters. Based on the mentioned parameters, the percentage of each parameter was computed using (1) (Hamzei and Navimipour, 2018). This equation means that the number of each occurrence was counted and divided by the sum of the whole number of occurrences, then the answer was multiplied by 100 (Eq. (1)).

$$\text{Percentage of occurrence (i)} = \frac{Number\ of\ each\ \text{occurrence}}{\sum Number\ of\ all\ occurrence} * 100 \tag{1}$$

Fig. 10 indicates that in topical learning approaches, researchers focused on accuracy (23%) and recall (15%), while in opinion/sentiment learning approaches, accuracy (31%) and F-measure (18%) are the crucial ones. The significant parameters in embedding learning approaches were time and cost by 23% and 16%, respectively. To say more, 20% of the papers with community learning approaches have optimized scalability and 18% of them have reduced time, so the results showed that accuracy is essential in most approaches; however, privacy, reliability, and security are somewhat neglected in these approaches.

- Q3: What are the tools used in social network analysis and big data areas?

Some of the papers did not mention any tools for analyzing and implementing the intended approaches. According to tool columns in Tables 3, 5, 7, and 9, along with python programming language, Hadoop was the top used tool in 74 research studies of social network analysis. The high frequent application of Hadoop is due to its open-source libraries for distributed and parallel processing of large datasets, cost-effective, big storage, reliability, scalability, and handling unstructured and semi-structured data.

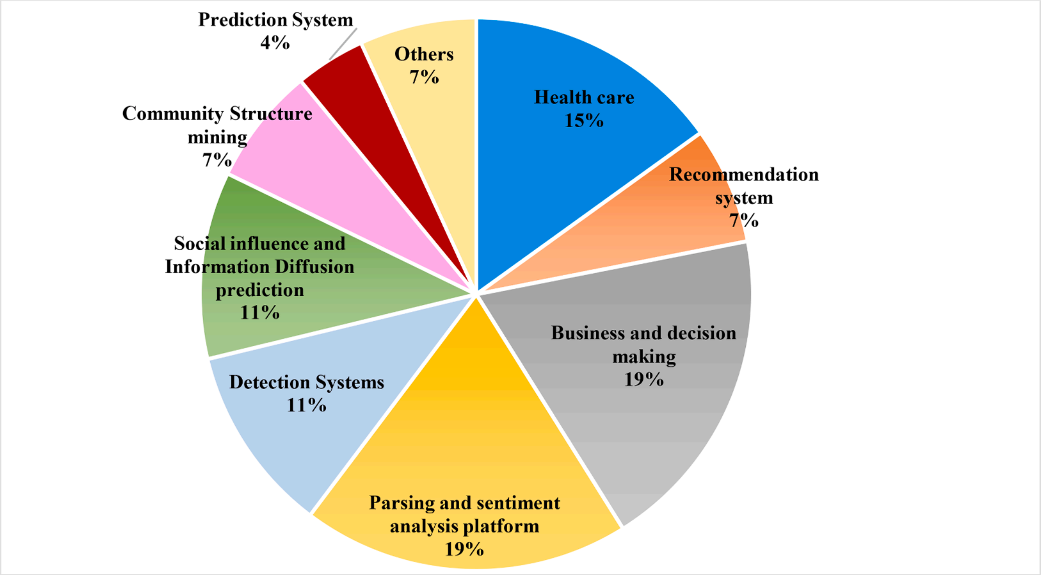

Fig. 11. Percentage of social big data analysis applications in the studied papers.

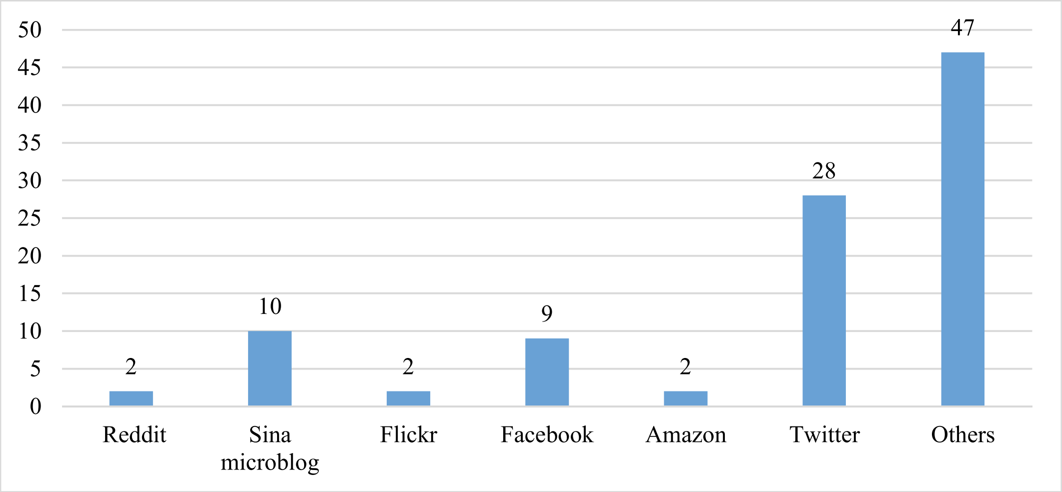

Fig. 12. Repetition of used datasets and case studies in the selected papers.

- Q4: What are the social big data analysis applications in the studied papers?

Fig. 11 demonstrates the social big data analysis applications of the reviewed papers, along with their percentage of repetitions. The results showed that, in the reviewed papers, the business and decision making, and parsing and sentiment analysis platform had the highest applications with 19% each. Along with these two applications, health care (15%) was a significant application of big social data analysis in studied papers.

- Q5: What are the datasets and case studies used in social big data analysis?

Selected studies have used various datasets to evaluate their approaches for analyzing the results of experiments. Based on the findings shown in Fig. 12, most of the researchers used Twitter. In addition to Twitter, the most significant percentage of the usage of datasets belongs to Sina microblog and Facebook.

- Q6: What evaluation methods are applied to measure the big data analytic approaches in social networks?

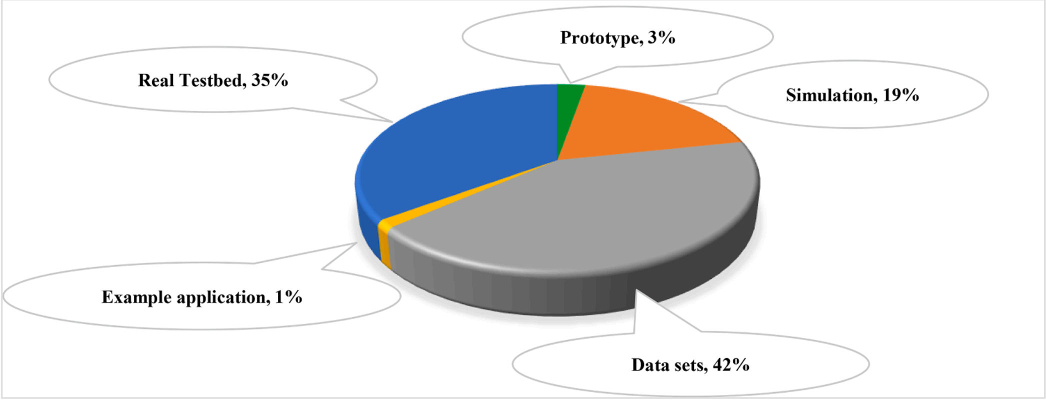

Fig. 13. Percentage of evaluation methods in the selected papers.

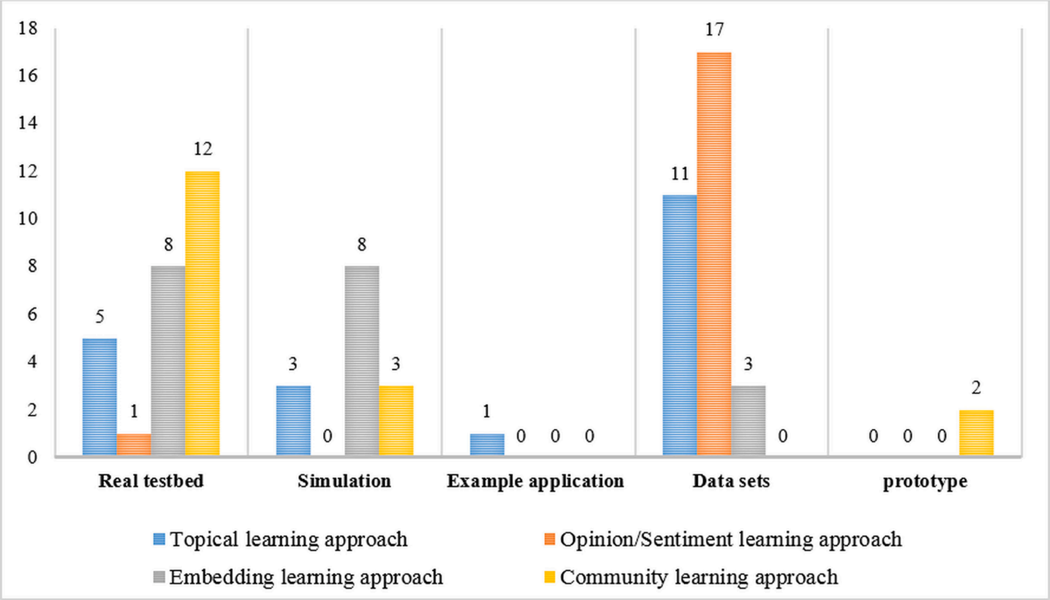

Fig. 14. Repetition of evaluation methods in each approach in the selected papers.

Based on Tables 3, 5, 7, and 9, which have depicted the evaluation methods applied in each approach, there were five evaluation methods in the reviewed papers: Simulation, prototype, data sets, real testbed, and example application. As shown in Fig. 13, 42% of assessments were related to data sets, while 35% of them were associated with real testbed. Lucidly, simulation dedicated 19% in itself. Fig. 14, displays the repetition of evaluation methods in each learning approach. The comparative results illustrate that in topical and opinion/sentiment learning, most evaluation methods are data sets. ML algorithms and data sets were widely used in semantic analysis and incorporated many ideas and innovations into social networks, welcoming virtual world users and social network growth; however, in community learning approaches, the real testbed has the highest usage in most evaluations. Finally, real testbed and simulation cover most of the evaluations for embedding learning approaches.

6. Open issues and future directions

Given the vast quantity of live social media streams and their impact on society, many techniques have been proposed to collect and analyse live UGC to support various applications. The techniques studied in this paper assist us in gaining insights into social data via big data analytics. The presented systematic literature is a good starting point to reveal open challenges. However, content-oriented and network-oriented approaches still face many vital challenges as mentioned below:

- Q7: What are the challenges and future perspectives of big data analytic approaches in social networks?

- The extensive usage of social media has resulted in the advancement of many disciplines and industries in which healthcare is one vertical application that has attracted much attention. Fig. 11 demonstrates an increasing tendency towards healthcare systems along with other domains. Patients join different social media groups, sharing experiences, describing their illness, and the treatment process. Social platforms provide patients with emotional supports from peers with similar conditions. The first-hand experiences and comments from other members in the network are invaluable sources for making informed decisions, especially for those with chronic conditions (Akbari et al., 2019, 2018). Further, healthcare professionals also utilize social media to share healthcare, psychology, and medical information and to interact with their peers as well as patients (Nie et al., 2014).

In this respect, public care organizations can start-up social health networks for diagnosing and preventing the spread of contagious disease in various geographical locations at different times by exploring public health posts in various social networks (Elkin et al., 2017). On the other hand, by analyzing the graph of interactions between users on social networks and examining influential users, nodes with multiple edges have been identified, so by limiting and quarantining them, the transmission rate of contagious disease can be forecasted, which allows us in better decision making to control infectious ailments. This would ultimately lead to a notable reduction in healthcare costs (Zadeh et al., 2019).

They can also track the origin of diseases, the transmission of diseases from generation to generation, the effects of drugs, and their interactions in different diseases (Thorstad and Wolff, 2019). This helps the pharmaceutical industry as well as healthcare promotion and health disorder diagnosis. One of the limitations of the current work in this area is that the nodes and their relations were considered static over time. Considering and analyzing the network in real-time and the dynamic interaction among nodes are still open issues that can achieve more accurate predictions. Most researchers also have studied social influence and information diffusion in a particular platform; analyzing information diffusion and social influence across multiple platforms simultaneously can also be a challenge in the future. However, among the reviewed literature, there were few papers on political and e-commerce applications, so these two issues are good topics for future research.

- In case of a vast number of data sources, another challenge is enhancing accuracy to improve services and predictions in various social network applications. For example, in social networking services, users frequently publish about themselves via status updates, photos, videos, self-description, and interests. Some of the recommendation and prediction systems predict the users' personalities by considering the users' profile data. On the other hand, some people keep some of their personal information private, or some users deliberately create fake accounts or fake information such as birth date, location, occupation, and status to increase the number of followers or get more likes; the available data may be fake or cannot be achieved due to privacy concerns, so the result of prediction is not accurate. Further, user profiling would be an essential aspect of social networking services to attest accurate prediction and recommendation (Akbari and Chua, 2017; Akbari et al., 2017).
- The ever-increasing volume of social media data has led to the distribution of files in various physical locations. A key future direction is to investigate factors such as network traffic, data locality, latency, high-level runtime of feature extraction, and clustering users. Despite the fact that enhancing the speed of feature extraction has been considered by a limited number of papers (Hsu et al., 2017), other challenges are still unsolved.
- Conspicuously, due to the high volume of data and the rapid growth of contents produced in social platforms, scalability is still a key factor to determine the effectiveness of social network analysis frameworks. The scalability issue includes handling an immense number of users, updating users' profiles and status, internal network traffic, as well as data storage and database management, so the expanding infrastructure, infrastructure management, and operational costs can affect the scalability challenge. Although some papers have proposed algorithms or methods to increase scalability in their approaches (Sachar and Khullar, 2017; Feng et al., 2018; Aa et al., 2015), others implement their approaches on small scale datasets; hence, it is still a significant challenge.
- The Internet has increased the growth of social networks to connect people and make it easier for them to find friends and share multimedia information, such as photos, videos, which are considered big data in social networks. With the increasing likelihood of cyber-attacks or malicious users, there is a risk of personal data being misused. A limited number of studies made efforts to solve this issue (Zhou et al., 2016); hence, offering novel approaches to ensure the privacy-preserving of social network users to secure photos, videos, sensitive personal data, and profiles, without crippling the utility of social media data, is a crucial challenge for future research.
- Due to the streaming nature of social data, both collecting and analyzing real-time data from various sources can assist the organization of customers' tweets, blog posts, and status updates. It allows organizations to track and answer customers' updates and comments as soon as possible. Some papers debated this challenge (Sayed et al., 2020; Lee and Paik, 2017; Rodrigues and Chiplunkar, 2019), but, unlike Twitter Streaming API, Facebook's graph API does not provide any real-time streaming access. The analytic approach should be able to investigate social media platforms in real-time, which leads to real-time results, so the real-time nature of social data is still an appealing challenge.
- Predictive analytics is another interesting direction that still remains as an open and challenging task. The key challenges which were focused on in (Yang et al., 2015) are aggregating data, extracting high dimensionality features, and building a model that can predict future events. Considerable amounts of data that are produced by users in social networks represent the views, suggestions, and thoughts of users in the form of texts, images, and videos, which may be high-quality or low-quality. As these data come from grass roots users with informal and unconstructed formats, social data are popular as noisy sources of information. Low quality, out of date, or incorrect data can lead to wrong or inaccurate analytics results; therefore, in addition to extracting high-quality information from a variety of sources, it is also essential to prevent the flow of misinformation. Although ML and data mining permit us to reduce the impact of low-quality data, it still cannot assure the proper quality of data. Besides, the modeling process should be

repeatable to ensure and extract meaningful relationships among data. Without a useful model, the predictive system cannot produce satisfactory results; therefore, data quality and modeling are two engrossing directions for future works in predictive analysis.

- From the sentiment analysis aspect, the following key challenges are still open to be addressed:

o *Domain dependency*: As sentiment analysis is a domain-dependent task in which the polarity of some words and phrases vary from one domain to the other; thus, a classifier trained for a specific domain may fail to perform well on other domains.
o *The rare-resource languages*: Most of the resources of sentiment analysis are only built for English language. There is no sufficient corpus for such languages as Chines, French, Hindi, Spanish, and so on. The bottleneck of performing opinion/sentiment analysis is the scarcity of predefined dictionaries and tools for various languages.
o *Detecting sarcasm*: Since sentiment analysis classifies texts as positive, negative, or neutral, another challenging issue in sentiment analysis is detecting sarcasm. It refers to sentences that have negative meanings despite the use of positive sentiment-bearing words. In other words, the meaning is just the opposite. It is a challenging task for a system to identify sarcastic sentences. Researchers should allocate their attention to find innovative approaches to analyse sarcasm in the sentiment of social big data analysis.
o *Detecting slang*: Most of the people use slang to express their feelings and, as slang words contain extreme sentiments, detecting slang words is a serious problem.
o *Heterogeneous nature of data*: The sentiment classifiers should work effectively and handle the diverse types of data from various data sources.
o *Unreliable and incomplete data*: Users usually use abbreviations in a social network. Social network data may contain a lot of noise and misspellings; the sentiment classification of these data is not accurate; therefore, sentiment classifiers should be able to predict incomplete information to have a more accurate prediction.
o *Semantic relations in multiple data sources*: Different social networks such as Twitter, Facebook, Instagram, and YouTube may discuss the same topic. Researchers in studied papers, investigate data only on a single social media, so the analysis of an event from various social media is a challenge that can offer better insights for the task of sentiment analysis and its model creation.
o *Subjectivity detection*: Regarding the personality of a user or his political views, a text may be neutral to one person, but not for the other, so a sentence may have a different interpretation.
o *Spam detection*: Spammers or fake users try to post fake reviews and to mislead other readers, so detecting these spams among posts is a significant challenge.

Many researchers try to mitigate a limited number of these challenges (Sun et al., 2018; Kauffmann et al., 2019; Jimenez-Marquez et al., 2019), but they failed to achieve high accuracy, so most of these challenges in sentiment analysis have not yet been resolved, and further research is needed.

- Finally, a few number of the studied papers did not test their approaches on real datasets of social networks. Unlike users' typical sentences, specialized sentences were not technically analysed. Also, specific vocabulary is used for particular platforms, e.g., the use of slang terms, which makes analysis very specific to each platform; therefore, it is another research direction, and further studies may test various social networks and real datasets of social networks. More experiments can be performed to increase the performance of social big data analytic approaches in the future.

7. Threats to validity and limitations

This SLR presents a taxonomy and a comparison of big data analytics in social networks. These types of review papers usually have constraints (Brereton et al., 2007), but the results of SLRs are mainly reliable (Zhang and Babar, 2013). The major limitations and threats to the validity of this SLR are discussed below.

- *The scope of the research:* In the paper selection process, only academic journals and conferences were considered. Furthermore, national conferences and journals, non-English papers, book chapters, and review papers were neglected.
- *Study and publication bias:* These nine electronic publishers offer the most related and valid papers; some of them were neglected via the paper selection process; therefore, the selection of all related papers cannot be guaranteed.
- *Study queries:* This paper is proposed according to seven questions, which were defined to find their answers. Other researchers may add some other questions.
- *Taxonomy:* The reviewed papers were classified into two main categories based on analysis methods: Content-oriented and network-oriented approaches, but it can be categorized otherwise.
- *Simulation:* The reviewed papers were not simulated.
- *Time range:* Only papers from 2013 to August 2020 were reviewed, and those before 2013 were not considered.

As a matter of fact, by defining a review protocol, following a systematic procedure, and the involvement of various researchers, this SLR has high validity.

8. Conclusion

This paper presents a systematic review of big data analytics in social networks. We explained the research methodology, paper selection process, and selected 74 papers between 2013 and August 2020, from among 785 papers in our search query. A significant number of the studied papers were related to IEEE, Springer, and ScienceDirect journals, with 37%, 27%, and 19%, respectively. On the other hand, each of Taylor&Francis and Emerald publishers with 1% had the lowest number of published papers. From these studies, 74 papers were categorized into two approaches: Content-oriented approaches (51%) and network-oriented approaches (49%). Besides, the main ideas, advantages, disadvantages, evaluation methods, tools, and evaluation parameters of each studied paper were discussed. It was found that the most widely considered evaluation parameters were accuracy (20%), time (16%), and scalability (12%), but privacy, reliability, and security measures were somewhat neglected. Considering the applied tools, it is observed that, in the selected studies, along with Python programming language, Hadoop was used more than other tools. Concerning the outcome of this SLR, the existing social big data analytic approaches have inadequate capability to guarantee privacy-preserving and scalability and have faced several open issues such as latency, real-time processing, and high run-time of feature selection. Lucidly, the most unresolved challenges are various aspects of opinion/sentiment analysis such as domain dependency, the rare resource languages, detecting sarcasm and slangs, subjectivity detection, and multiple data sources. We hope that the findings of this paper will assist researchers to propose novel contributions to overcome social big data challenges.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We are grateful for the insightful and constructive comments offered by Dr. Mohammad Akbari, and also appreciate anonymous reviewers for their precious comments which improved the final version.

Appendix A. List of evaluation parameters and their description

Evaluation parameter	Description and formula
Confusion matrix	For a binary classifier, four possible outputs of the confusion matrix are defined as below: ***True Positive*** **(TP):** The number of correctly positive predictions ***True Negative*** **(TN):** The number of correctly negative predictions ***False Positive*** **(FP):** The number of predictions that are labelled positive incorrectly ***False Negative*** **(FN):** The number of predictions that are labelled negative incorrectly
Accuracy	Accuracy in the social network refers to the degree of similarity between the actual structure of a relationship and the individuals' perceptions of the structure of the same relationship in a particular social media (Casciaro et al., 1999). In other words, it is the number of correctly predicted observations over the total number of observations. $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
Precision	Precision focuses on false positives and is the number of correctly predicted positive observations over the total predicted positive observations. Indeed, it is the ability of the model not to label a negative sample as a positive. $Precision = \frac{TP}{TP+FP}$
Recall (Sensitivity or TPR)	Recall is the fraction of correctly predicted positive observations by a proposed model among all positive observations in the actual class of the dataset. Intuitively, it is the ability of a classifier to discover all positive samples correctly. $Recall = \frac{TP}{TP+FN}$
F-measure	F-measure is a harmonic mean of precision and recall to identify if a presented model reaches the objective of high precision and recall at a time. Since it is a weighted average of precision and recall and takes both FP and FN into account, it can be applied for measuring the efficiency of the model in many domains. $F-measure = 2*\frac{Precision*Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN}$
Specificity (TNR)	Specificity shows what the proportion of negative observations is predicted correctly. $Specificity = \frac{TN}{TN+FP}$
ROC (AUC)	ROC curve is illustrated graphically to show the trade-off between sensitivity (TPR) on Y-axis and (1-specificity) (FPR) on X-axis for every possible threshold value. The area under the curve refers to AUC that is applied to determine the ability of a classifier in distinguishing positive and negative classes. The higher the AUC, the better the performance of a classifier is.
Kappa coefficient	Kappa is an inter-rater reliability measure to evaluate the agreement between two raters. In other words, it shows how closely the observations classified by a classifier are in agreement with the data labeled as ground truth. It can be calculated by this formula: $Kappa = (observedaccuracy - expectedaccuracy)(1 - expectedaccuracy)$
Matthews Correlation Coefficient (MCC)	It evaluates the correlation between the observed and predicted classifications of an instance. The formula of the MCC is:

(continued on next page)

(*continued*)

Evaluation parameter	Description and formula
	$$\boldsymbol{MCC} = \frac{(\boldsymbol{TP}*\boldsymbol{TN}) - (\boldsymbol{FP}*\boldsymbol{FN})}{\sqrt{(\boldsymbol{TP}+\boldsymbol{FP})(\boldsymbol{TP}+\boldsymbol{FN})(\boldsymbol{TN}+\boldsymbol{FP})(\boldsymbol{TN}+\boldsymbol{FN})}}$$
Clustering coefficient	Clustering coefficient indicates how much each node is willing to create clusters in a network. There are two types of clustering coefficients: the local clustering and the global clustering. The local refers to the embeddedness of every single node, while the global refers to an overall indication of clustering in the network. A clustering parameter is a real number between zero and one (Kalna and Higham, 2007; Zhang et al., 2008). When there are no clusters, this coefficient is equal to zero, and in case of disjoint cliques, in which the maximal clustering occurs, it is equal to one (Holland and Leinhardt, 1971; Watts and Strogatz, 1998).
Security	Security refers to the requirements that the system needs to protect against potential attacks, threats, unauthorized access, and privacy-preserving issues (Cutillo et al., 2010).
Scalability	Scalability means the ability of a social network to expand in case of rising demand for processors, networks, or file system resources. Scalability consists of two categories, as follows: o ***Scale horizontally (or scale out/in):*** It refers to the addition of new hardware instead of increasing the capability of the existing hardware. o ***Scale vertically (or scale up/down):*** It can be performed by adding resources, or powerful hardware to (or removing resources from) a system like adding CPU or RAM to a single system node or a single computer.
Time	In this paper, all the factors related to time, such as execution time, average response time, statistical analysis time (starting time), delay, and running time are considered as the time factor.
Normalized mutual information (NMI)	NMI is an information theoretic-based measure that can be used to assess the quality of clustering to compare community detection methods. This measure compares different clusters, and whenever its value is high, it means that the two clusters are similar (Amelio and Pizzuti, 2017). If clusters X and Y are precisely the same, their NMI is equal to one (Wang et al., 2010).
Cost	The price of acquiring, producing, performing, or maintaining the requested service
Influence diffusion	This measure shows how one person's actions affect other people in a network (Junquero-Trabado et al., 2011); it shows how many users are affected by the most influential users in the network.
Centrality measures	In the context of web information retrieval, using centrality measures is a vital task in community analysis (Getoor and Diehl, 2005). By using centrality measures, researchers try to answer the question "who is the most important, impressive, or central person in the network?" (Abbasi et al., 2011). Some of the popular centrality measures are discussed below: o ***Degree centrality*:** It means the degree and the number of neighbors of a node and is computed by the number of direct links to a node. In the undirected graph, the more central the node is, the higher the degree will be (Everett, 2016). In a digraph, there are two types of this measure, in-degree, which refers to the number of inbound links to a node, and out-degree, which is the number of outbound links of a node (Kim et al., 2011). o ***Closeness centrality*:** Closeness centrality that calculates the shortest path among all nodes and is defined for a node V as the inverse of the distance (Eq. (1)). In other words, closeness means the length of time it takes to transfer information from one node to all other nodes (Luenberger, 1979). $$\boldsymbol{closeness}(\boldsymbol{v}) = \frac{1}{\sum_{i\neq v}\boldsymbol{d}_{vi}}$$ o ***Betweenness centrality*:** It refers to the number of times a node is placed among the shortest paths of other nodes, that is, after identifying all the shortest paths, the number of paths in which a given node is located is counted (Newman, 2004). o ***Eigenvector centrality*:** Eigenvector centrality is different from in-degree centrality, referring to the importance of each node of the graph. A node with high in-degree centrality does not necessarily have a high eigenvector centrality and vice versa (Catanese et al., 2011), so this parameter shows the important nodes that influence the entire network (Newman, 2004). o ***PageRank*:** PageRank is calculated to determine the importance of the node by considering the degree and quality of the nodes. It focuses on the centrality of linkers, link directions, and their weights (Newman, 2004). It is a recursive measure where the value for one node grows with the PageRank of its neighbors weighted by the reciprocal of their degrees. It can be thought of as the probability of visiting a node under the random surfer model (Page et al., 1999).

References

Arora, A., Bansal, S., Kandpal, C., Aswani, R., Dwivedi, Y., 2019. Measuring social media influencer index-insights from facebook, Twitter and Instagram. J. Retail. Cons. Serv. 49, 86–101.

Lai, W.K., Chen, Y.U., Wu, T.-Y., 2020. Analysis and evaluation of random-based message propagation models on the social networks. Comput. Netw. 170, 107047.

Alalwan, A.A., Rana, N.P., Dwivedi, Y.K., Algharabat, R., 2017. Social media in marketing: A review and analysis of the existing literature. Telematics Inform. 34 (7), 1177–1190.

R. Kumar, J. Novak, and A. Tomkins, Structure and evolution of online social networks. In Link mining: models, algorithms, and applications: Springer, 2010, pp. 337–357.

Feng, Y., Zhou, P., Wu, D., Hu, Y., 2018. Accurate content push for content-centric social networks: A big data support online learning approach. IEEE Trans. Emerg. Top. Comput. Intell. 99, 1–13.

Heidemann, J., Klier, M., Probst, F., 2012. Online social networks: A survey of a global phenomenon. Comput. Netw. 56 (18), 3866–3878.

Busalim, A.H., 2016. Understanding social commerce: A systematic literature review and directions for further research. Int. J. Inf. Manage. 36 (6), 1075–1088.

Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: Recent achievements and new challenges. Inf. Fusion 28, 45–59.

M. Jamali and H. Abolhassani, Different aspects of social network analysis. In Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on, 2006, pp. 66–72: IEEE.

Martinez-Rojas, M., del Carmen Pardo-Ferreira, M., Rubio-Romero, J.C., 2018. Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. Int. J. Inf. Manage. 43, 196–208.

Cetto, A., Klier, M., Richter, A., Zolitschka, J.F., 2018. "Thanks for sharing"—Identifying users' roles based on knowledge contribution in Enterprise Social Networks. Comput. Netw. 135, 275–288.

Go, E., You, K.H., 2016. But not all social media are the same: Analyzing organizations' social media usage patterns. Telematics Inform. 33 (1), 176–186.

[13] L. Manovich, Trending: The promises and the challenges of big social data. In Debates in the digital humanities, vol. 2, pp. 460–475, 2011.

Lomborg, S., Bechmann, A., 2014. Using APIs for data collection on social media. Inf. Soc. 30 (4), 256–265.

F. B. Abdesslem, I. Parris, and T. Henderson, Reliable online social network data collection. In Computational Social Networks: Springer, 2012, pp. 183–210.

Otte, E., Rousseau, R., 2002. Social network analysis: a powerful strategy, also for the information sciences. J. Inf. Sci. 28 (6), 441–453.

Cross, R., Borgatti, S.P., Parker, A., 2002. Making invisible work visible: Using social network analysis to support strategic collaboration. Calif. Manage. Rev. 44 (2), 25–46.

Parveen, F., Jaafar, N.I., Ainin, S., 2015. Social media usage and organizational performance: Reflections of Malaysian social media managers. Telematics Inform. 32 (1), 67–78.

Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15 (5), 662–679.

A. Katal, M. Wazid, and R. Goudar, Big data: Issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on, 2013, pp. 404–409: IEEE.

Terrazas, G., Ferry, N., Ratchev, S., 2019. A cloud-based framework for shop floor big data management and elastic computing analytics. Comput. Ind. 109, 204–214.

Canito, J., Ramos, P., Moro, S., Rita, P., 2018. Unfolding the relations between companies and technologies under the Big Data umbrella. Comput. Ind. 99, 1–8.

di Bella, E., Leporatti, L., Maggino, F., 2018. Big data and social indicators: Actual trends and new perspectives. Soc. Indic. Res. 135 (3), 869–878.

Hadi, M.S., Lawey, A.Q., El-Gorashi, T.E., Elmirghani, J.M., 2018. Big data analytics for wireless and wired network design: A survey. Comput. Netw. 132, 180–199.

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manage. 35 (2), 137–144.

Kitchin, R., 2014. The real-time city? Big data and smart urbanism. GeoJournal 79 (1), 1–14.

S. Sagiroglu and D. Sinanc, Big data: A review. In Collaboration Technologies and Systems (CTS), 2013 International Conference on, 2013, pp. 42–47: IEEE.

Pei, F.-Q., Li, D.-B., Tong, Y.-F., 2018. Double-layered big data analytics architecture for solar cells series welding machine. Comput. Ind. 97, 17–23.

Peng, S., Wang, G., Zhou, Y., Wan, C., Wang, C., Yu, S., 2017. An immunization framework for social networks through big data based influence modeling. IEEE Trans. Dependable Secure Comput.

Duan, Y., Edwards, J.S., Dwivedi, Y.K., 2019. Artificial intelligence for decision making in the era of Big Data–Evolution, challenges and research agenda. Int. J. Inf. Manage. 48, 63–71.

Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M., 2007. Lessons from applying the systematic literature review process within the software engineering domain. J. Syst. Softw. 80 (4), 571–583.

B. Kitchenham and S. Charters, Guidelines for performing systematic literature reviews in software engineering, 2007.

Jamshidi, P., Ahmad, A., Pahl, C., 2013. Cloud migration research: A systematic review. IEEE Trans. Cloud Comput. 1 (2), 142–157.

Jatoth, C., Gangadharan, G., Buyya, R., 2015. Computational intelligence based QoS-aware web service composition: A systematic literature review. IEEE Trans. Serv. Comput. 10 (3), 475–492.

Yaqoob, I., et al., 2016. TEMPORARY REMOVAL: Information fusion in social big data: Foundations, state-of-the-art, applications, challenges, and future research directions. Int. J. Inf. Manage.

Ghani, N.A., Hamid, S., Hashem, I.A.T., Ahmed, E., 2018. Social media big data analytics: A survey. Comput. Hum. Behav.

Bukovina, J., 2016. Social media big data and capital markets—An overview. J. Behav. Exp. Finance 11, 18–26.

M. E. Martin and N. Schuurman, Social media big data acquisition and analysis for qualitative GIScience: challenges and opportunities. Ann. Am. Assoc. Geogr., pp. 1–18, 2019.

M. Arnaboldi, C. Busco, and S. Cuganesan, Accounting, accountability, social media and big data: revolution or hype? Acc. Audit. Account. J., 2017.

Peng, S., Wang, G., Xie, D., 2016. Social influence analysis in social networking big data: Opportunities and challenges. IEEE Netw. 31 (1), 11–17.

I. Guellil and K. Boukhalfa, Social big data mining: A survey focused on opinion mining and sentiments analysis. In 2015 12th International Symposium on Programming and Systems (ISPS), 2015, pp. 1–10: IEEE.

S. Gole and B. Tidke, A survey of big data in social media using data mining techniques. In 2015 International Conference on Advanced Computing and Communication Systems, 2015, pp. 1–6: IEEE.

P. V. Paul, K. Monica, and M. Trishanka, A survey on big data analytics using social media data. In 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), 2017, pp. 1–4: IEEE.

Sebei, H., Taieb, M.A.H., Aouicha, M.B., 2018. Review of social media analytics process and Big Data pipeline. Social Netw. Anal. Min. 8 (1), 30.

Al-Garadi, M.A., et al., 2019. Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. IEEE Access 7, 70701–70718.

O. Lerena, F. Barletta, F. Fiorentin, D. Suárez, and G. Yoguel, Big data of innovation literature at the firm level: a review based on social network and text mining techniques. Econ. Innov. New Technol., pp. 1–17, 2019.

Kitchenham, B., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., Linkman, S., 2009. Systematic literature reviews in software engineering–A systematic literature review. Inf. Softw. Technol. 51 (1), 7–15.

Rahimi, M., Songhorabadi, M., Kashani, M.H., 2020. Fog-based smart homes: A systematic review. J. Netw. Comput. Appl., 102531

Haghi Kashani, M., Rahmani, A.M., Jafari Navimipour, N., 2020. Quality of service-aware approaches in fog computing. Int. J. Commun. Syst., e4340

C. Calero, M. F. Bertoa, and M. Á. Moraga, A systematic literature review for software sustainability measures. In 2013 2nd international workshop on green and sustainable software (GREENS), 2013, pp. 46–53: IEEE.

Aznoli, F., Navimipour, N.J., 2017. Deployment strategies in the wireless sensor networks: systematic literature review, classification, and current trends. Wireless Pers. Commun. 95 (2), 819–846.

Yang, M., Kiang, M., Shang, W., 2015. Filtering big data from social media–Building an early warning system for adverse drug reactions. J. Biomed. Inform. 54, 230–240.

Aa, V., Shekhara, V.S., Jb, R., Aggrawalb, T., Balasubramanya, K., Murthya, S.N., 2015. Cloud based big data analytics framework for face recognition in social networks using machine learning. Procedia Comput. Sci. 50, 623–630.

Moessner, M., Feldhege, J., Wolf, M., Bauer, S., 2018. Analyzing big data in social media: Text and network analyses of an eating disorder forum. Int. J. Eat. Disord.

Cheung, M., She, J., Jie, Z., 2015. Connection discovery using big data of user-shared images in social media. IEEE Trans. Multimedia 17 (9), 1417–1428.

N. Straton, R. R. Mukkamala, and R. Vatrapu, Big social data analytics for public health: Predicting facebook post performance using artificial neural networks and deep learning. In 2017 IEEE International Congress on Big Data (BigData Congress), 2017, pp. 89–96: IEEE.

P. Sachar and V. Khullar, Social media generated big data clustering using genetic algorithm. In 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017, pp. 1–6: IEEE.

A. Vakali, N. Kitmeridis, and M. Panourgia, A distributed framework for early trending topics detection on big social networks data threads. In INNS Conference on Big Data, 2016, pp. 186–194: Springer.

Huo, Y., Ma, L., Zhong, Y., 2018. A Big Data privacy respecting dissemination method for social network. J. Signal Process. Syst. 90 (4), 467–475.

A. H. Zadeh, H. M. Zolbanin, R. Sharda, and D. Delen, Social media for nowcasting flu activity: Spatio-temporal big data analysis. Inf. Syst. Front., pp. 1–18, 2019.

Xylogiannopoulos, K.F., Karampelas, P., Alhajj, R., 2020. A password creation and validation system for social media platforms based on big data analytics. J. Ambient Intell. Hum. Comput. 11 (1), 53–73.

Subroto, A., Apriyana, A., 2019. Cyber risk prediction through social media big data analytics and statistical machine learning. J. Big Data 6 (1), 50.

D. Makaroğlu, A. Çakır, and K. Kocabaş, Social Media and Clickstream Analysis in Turkish News with Apache Spark. In International Conference on Intelligent and Fuzzy Systems, 2019, pp. 221–228: Springer.

Singh, A., Kaur, M., 2019. Intelligent content-based cybercrime detection in online social networks using cuckoo search metaheuristic approach. J. Supercomput. 1–23.

R. Thorstad and P. Wolff, Predicting future mental illness from social media: A big-data approach. Behav. Res. Methods, pp. 1–15, 2019.

E. Alomari, I. Katib, and R. Mehmood, Iktishaf: A Big Data road-traffic event detection tool using twitter and spark machine learning. Mob. Netw. Appl., pp. 1–16, 2020.

Panarello, A., Celesti, A., Fazio, M., Puliafito, A., Villari, M., 2020. A big video data transcoding service for social media over federated clouds. Multimedia Tools Appl. 79 (13), 9037–9061.

Sahoo, S.R., Gupta, B., 2020. Fake profile detection in multimedia big data on online social networks. Int. J. Inf. Comput. Secur. 12 (2–3), 303–331.

Zhou, P., Zhou, Y., Wu, D., Jin, H., 2016. Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks. IEEE Trans. Multimedia 18 (6), 1217–1229.

Zhang, C., Xie, L., Aizezi, Y., Gu, X., 2019. User multi-modal emotional intelligence analysis method based on deep learning in social network Big Data environment. IEEE Access 7, 181758–181766.

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., Mora, H., 2019. A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. Ind. Mark. Manage.

Jiang, D., Luo, X., Xuan, J., Xu, Z., 2017. Sentiment computing for the news event based on the social media big data. IEEE Access 5, 2373–2382.

Dalla Valle, L., Kenett, R., 2018. Social media big data integration: A new approach based on calibration. Expert Syst. Appl. 111, 76–90.

Jimenez-Marquez, J.L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J.L., Ruiz-Mezcua, B., 2019. Towards a big data framework for analyzing social media content. Int. J. Inf. Manage. 44, 1–12.

Shirdastian, H., Laroche, M., Richard, M.-O., 2019. Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. Int. J. Inf. Manage. 48, 291–307.

Zhu, B., Zheng, X., Liu, H., Li, J., Wang, P., 2020. Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. Chaos, Solitons Fractals 140, 110123.

Fan, M., Billings, A., Zhu, X., Yu, P., 2020. Twitter-based BIRGing: Big Data analysis of English national team fans during the 2018 FIFA World Cup. Commun. Sport 8 (3), 317–345.

C. Lee and I. Paik, Stock market analysis from Twitter and news based on streaming big data infrastructure. In 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), 2017, pp. 312–317: IEEE.

A. A. Sayed, M. M. Abdallah, A. M. Zaki, and A. A. Ahmed, Big Data analysis using a metaheuristic algorithm: Twitter as Case Study. In 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), 2020, pp. 20–26: IEEE.

van Dieijen, M., Borah, A., Tellis, G.J., Franses, P.H., 2020. Big data analysis of volatility spillovers of brands across social media and stock markets. Ind. Mark. Manage. 88, 465–484.

Spruce, M., Arthur, R., Williams, H., 2020. Using social media to measure impacts of named storm events in the United Kingdom and Ireland. Meteorol. Appl. 27 (1), e1887.

Um, J.-H., Jeong, C.-H., Choi, S.-P., Lee, S., Kim, H.-M., Jung, H., 2013. Distributed and parallel big textual data parsing for social sensor network. Int. J. Distrib. Sens. Netw. 9 (12), 525687.

I. Moise, The technical hashtag in Twitter data: A hadoop experience. In 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 3519–3528: IEEE.

D. Hsu, M. Moh, and T.-S. Moh, Mining frequency of drug side effects over a large twitter dataset using apache spark. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 2017, pp. 915–924.

A. Baltas, A. Kanavos, and A. K. Tsakalidis, An apache spark implementation for sentiment analysis on twitter data. In International Workshop of Algorithmic Aspects of Cloud Computing, 2016, pp. 15–25: Springer.

X. Sun, C. Zhang, S. Ding, and C. Quan, Detecting anomalous emotion through big data from social networks based on a deep learning method. Multimedia Tools Appl., pp. 1–22, 2018.

BalaAnand, M., Karthikeyan, N., Karthik, S., 2019. Envisioning social media information for big data using big vision schemes in wireless environment. Wireless Pers. Commun. 1–20.

A. P. Rodrigues and N. N. Chiplunkar, A new big data approach for topic classification and sentiment analysis of Twitter data. Evol. Intell., pp. 1–11, 2019.

Persico, V., Pescapé, A., Picariello, A., Sperlí, G., 2018. Benchmarking big data architectures for social networks data processing using public cloud platforms. Future Gener. Comput. Syst. 89, 98–109.

Elkin, L.S., Topal, K., Bebek, G., 2017. Network based model of social media big data predicts contagious disease diffusion. Inf. Disc. Del. 45 (3), 110–120.

Gao, S., Pang, H., Gallinari, P., Guo, J., Kato, N., 2017. A novel embedding method for information diffusion prediction in social network big data. IEEE Trans. Ind. Inf. 13 (4), 2097–2105.

A. Talukder and C. S. Hong, A heuristic mixed model for viral marketing cost minimization in social networks. In 2019 International Conference on Information Networking (ICOIN), 2019, pp. 141–146: IEEE.

Chen, S., Yin, X., Cao, Q., Li, Q., Long, H., 2020. Targeted influence maximization based on cloud computing over big data in social networks. IEEE Access 8, 45512–45522.

Y. Wang, B. Zhang, A. V. Vasilakos, and J. Ma, PRDiscount: A heuristic scheme of initial seeds selection for diffusion maximization in social networks. In International Conference on Intelligent Computing, 2014, pp. 149–161: Springer.

Kumaran, P., Chitrakala, S., 2017. Social influence determination on big data streams in an online social network. Multimedia Tools Appl. 76 (21), 22133–22167.

Wu, Y., Huang, H., Wu, N., Wang, Y., Bhuiyan, M.Z.A., Wang, T., 2020. An incentive-based protection and recovery strategy for secure big data in social networks. Inf. Sci. 508, 79–91.

Wu, Y., Huang, H., Zhao, J., Wang, C., Wang, T., 2018. Using mobile nodes to control rumors in big data based on a new rumor propagation model in vehicular social networks. IEEE Access 6, 62612–62621.

Wu, J., Zhao, M., Chen, Z., 2018. Small data: Effective data based on big communication research in social networks. Wireless Pers. Commun. 99 (3), 1391–1404.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., Baesens, B., 2019. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Appl. Soft Comput. 74, 26–39.

Yang, X., McEwen, R., Ong, L.R., Zihayat, M., 2020. A big data analytics framework for detecting user-level depression from social networks. Int. J. Inf. Manage. 54, 102141.

Raj, E.D., Babu, L.D., 2015. A firefly swarm approach for establishing new connections in social networks based on big data analytics. Int. J. Commun. Netw. Distrib. Syst. 15 (2–3), 130–148.

K. Xu, F. Wang, X. Jia, and H. Wang, The impact of sampling on big data analysis of social media: A case study on flu and ebola. In 2015 IEEE Global Communications Conference (GLOBECOM), 2015, pp. 1–6: IEEE.

Su, Z., Xu, Q., Qi, Q., 2016. Big data in mobile social networks: A QoE-oriented framework. IEEE Network 30 (1), 52–57.

K. S. Kumar, D. E. Geetha, N. Nagesh, and T. S. Manoj, Identify the influential user in online social networks using R, Hadoop and Python. In 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), 2016, pp. 1–6: IEEE.

Y. Zhang, Z. Huang, S. Wang, X. Wang, and T. Jiang, “Spark-based measurement and analysis on offline mobile application market over device-to-device sharing in mobile social networks. in 2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS), 2017, pp. 545–552: IEEE.

Maireder, A., Weeks, B.E., Gil de Zúñiga, H., Schlögl, S., 2017. Big data and political social Networks: Introducing audience diversity and communication connector bridging measures in social network theory. Social Sci. Comput. Rev. 35 (1), 126–141.

Dabas, C., 2017. Big data analytics for exploratory social network analysis. Int. J. Inf. Technol. Manage. 16 (4), 348–359.

H. Aksu, M. Canim, Y.-C. Chang, I. Korpeoglu, and Ö. Ulusoy, Multi-resolution social network community identification and maintenance on big data platform. In Big Data (BigData Congress), 2013 IEEE International Congress on, 2013, pp. 102–109: IEEE.

Z. Wu, J. Chen, and Y. Zhang, An incremental community detection method in social big data. In 2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 2018, pp. 136–141: IEEE.

S. Yousfi, D. Chiadmi, F. Nafis, Toward a Big Data-as-a-service for social networks graphs analysis. In Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015, 2016, pp. 593–598: Springer.

Sun, J., Xu, W., Ma, J., Sun, J., 2015. Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework. Int. J. Prod. Econ. 165, 185–193.

Ghosh, G., Banerjee, S., Yen, N.Y., 2016. State transition in communication under social network: An analysis using fuzzy logic and density based clustering towards big data paradigm. Future Gener. Comput. Syst. 65, 207–220.

Wang, F., Mack, E.A., Maciewjewski, R., 2017. Analyzing entrepreneurial social networks with big data. Ann. Am. Assoc. Geogr. 107 (1), 130–150.

K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, Localization based on social big data analysis in the vehicular networks. IEEE Trans. Ind. Inform, 99(1), 2016.

C. Li, P. Zhou, Y. Zhou, K. Bian, T. Jiang, and S. Rahardja, Distributed private online learning for social big data computing over data center networks. In 2016 IEEE International Conference on Communications (ICC), 2016, pp. 1–6: IEEE.

I. Paik, Y. Koshiba, and T. A. S. Siriweera, Efficient service discovery using social service network based on big data infrastructure. In 2017 IEEE 11th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), 2017, pp. 166–173: IEEE.

J. Wang, C. Jiang, S. Guan, L. Xu, and Y. Ren, Big data driven similarity based U-model for online social networks. In GLOBECOM 2017-2017 IEEE Global Communications Conference, 2017, pp. 1–6: IEEE.

S. Sharma, Building Real-time knowledge in Social Media on Focus Point: An Apache Spark Streaming Implementation. In 2018 IEEE Punecon, pp. 1–6: IEEE.

H. F. Karimi, S. U. Masruroh, F. Mintarsih, The influence of iteration calculation manipulation on social network analysis toward twitter's users against hoax in Indonesia with single cluster multi-node method using apache Hadoop Hortonworkstm distribution. In 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018, pp. 1–6: IEEE.

W. Du, Toward semantic social network analysis for business big data. In 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), 2018, pp. 1–8: IEEE.

C. K. Leung and H. Zhang, Management of distributed big data for social networks. In 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2016, pp. 639–648: IEEE.

Jin, S., Lin, W., Yin, H., Yang, S., Li, A., Deng, B., 2015. Community structure mining in big data social media networks with MapReduce. Cluster computing 18 (3), 999–1010.

Kuang, L., Tang, X., Yu, M., Huang, Y., Guo, K., 2016. A comprehensive ranking model for tweets big data in online social network. EURASIP J. Wire. Commun. Netw. 2016 (1), 46.

Hamzei, M., Navimipour, N.J., 2018. Toward efficient service composition techniques in the Internet of things. IEEE Internet Things J. 5 (5), 3774–3787.

M. Akbari, X. Hu, and T.-S. Chua, Learning wellness profiles of users on social networks: The case of diabetes. In Social Web and Health Research: Springer, 2019, pp. 139–169.

M. Akbari, K. Relia, A. Elghafari, R. Chunara, From the user to the medium: Neural profiling across web communities. In Twelfth International AAAI Conference on Web and Social Media, 2018.

Nie, L., Zhao, Y.-L., Akbari, M., Shen, J., Chua, T.-S., 2014. Bridging the vocabulary gap between health seekers and healthcare knowledge. IEEE Trans. Knowl. Data Eng. 27 (2), 396–409.

M. Akbari and T.-S. Chua, Leveraging behavioral factorization and prior knowledge for community discovery and profiling. Presented at the Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, United Kingdom, 2017.

Akbari, M., Hu, X., Wang, F., Chua, T., 2017. Wellness representation of users in social media: Towards joint modelling of heterogeneity and temporality. IEEE Trans. Knowl. Data Eng. 29 (10), 2360–2373.

Zhang, H., Babar, M.A., 2013. Systematic reviews in software engineering: An empirical investigation. Inf. Softw. Technol. 55 (7), 1341–1354.

Casciaro, T., Carley, K.M., Krackhardt, D., 1999. Positive affectivity and accuracy in social network perception. Motiv. Emotion 23 (4), 285–306.

Kalna, G., Higham, D.J., 2007. A clustering coefficient for weighted networks, with application to gene expression data. AI Commun. 20 (4), 263–271.

Zhang, P., Wang, J., Li, X., Li, M., Di, Z., Fan, Y., 2008. Clustering coefficient and community structure of bipartite networks. Physica A 387 (27), 6869–6875.

Holland, P.W., Leinhardt, S., 1971. Transitivity in structural models of small groups. Comp. Group Stud. 2 (2), 107–124.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world'networks. Nature 393 (6684), 440.

L. A. Cutillo, M. Manulis, T. Strufe, Security and privacy in online social networks. In Handbook of Social Network Technologies and Applications .Springer, 2010, pp. 497–522.

Amelio, A., Pizzuti, C., 2017. Correction for closeness: Adjusting normalized mutual information measure for clustering comparison. Comput. Intell. 33 (3), 579–601.

X. Wang, L. Tang, H. Gao, H. Liu. Discovering overlapping groups in social media. In Data Mining (ICDM), 2010 IEEE 10th International Conference on, 2010, pp. 569–578: IEEE.

V. Junquero-Trabado, N. Trench-Ribes, M. A. Aguila-Lorente, D. Dominguez-Sal, Comparison of influence metrics in information diffusion networks. In Computational Aspects of Social Networks (CASoN), 2011 International Conference on, 2011, pp. 31–36: IEEE.

Getoor, L., Diehl, C.P., 2005. Link mining: A survey. Acm Sigkdd Explor. News. 7 (2), 3–12.

Abbasi, A., Altmann, J., Hossain, L., 2011. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. J. Inf. 5 (4), 594–607.

Everett, M.G., 2016. Centrality and the dual-projection approach for two-mode social network data. Methodol. Innovations 9.

Kim, Y., Choi, T.Y., Yan, T., Dooley, K., 2011. Structural investigation of supply networks: A social network analysis approach. J. Oper. Manage. 29 (3), 194–211.

D. G. Luenberger, Introduction to Dynamic Systems: Theory, Models, and Applications. Wiley New York, 1979.

Newman, M.E., 2004. Analysis of weighted networks. Phys. Rev. E 70 (5), 056131.

S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Crawling facebook for social network analysis purposes. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 2011, p. 52: ACM.

L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bringing order to the web. Stanford InfoLab1999.