A Systematic Review towards Big Data Analytics in Social Media

Ingole Sheetal Prakash, Student SEM-II, M.Tech-Computer Science (Artificial Intelligence), NIILM University,Kaithal,Haryana, India

Abstract

The recent advancement in internet 2.0 creates a scope to connect people worldwide using society 2.0 and web 2.0 technologies. This new era allows the consumer to directly connect with other individuals, business corporations, and the government. People are open to sharing opinions, views, and ideas on any topic in different formats out loud. This creates the opportunity to make the "Big Social Data" handy by implementing machine learning approaches and social data analytics. This study offers an overview of recent works in social media, data science, and machine learning to gain a wide perspective on social media big data analytics. We explain why social media data are significant elements of the improved data-driven decision-making process. We propose and build the "Sunflower Model of Big Data" to define big data and bring it up to date with technology by combining 5 V's and 10 Bigs. We discover the top ten social data analytics to work in the domain of social media platforms. A comprehensive list of relevant statistical/machine learning methods to implement each of these big data analytics is discussed in this work. "Text Analytics" is the most used analytics in social data analysis to date. We create a taxonomy on social media analytics to meet the need and provide a clear understanding. Tools, techniques, and supporting data type are also discussed in this research work. As a result, researchers will have an easier time deciding which social data analytics would best suit their needs.

Key words: big data; social media; big data analytics; social media analytics; text analytics; image analytics; audio analytics; video analytics; predictive analytics; descriptive analytics; prescriptive analytics; diagnostic analytics

1 Introduction

Big data have become a valuable resource. The utilization of big data is everywhere, starting from social networks, academia, healthcare, aerospace, transport planning, oil and gas development to telecoms, e-commerce, finance and insurance, military and surveillance, and a variety of other fields[1]. However, this massive amount of data will become an asset only if we know how to make data talk. Data analytics is that tool that makes data convey stories in a clear and accessible manner.

A well-known social listening platform—Brandwatch, says that over 3.4 billion people worldwide were using social media sites in May 2019[2]. A massive amount of structured, semi-structured, and unstructured data are produced in a short period of time in this social media based platform[3–6]. The reason for this is that a social media based platform allows for faster information sharing, supports text, image, audio, and video sharing, and allows the individual user to communicate with a huge number of other users at the same time. Surprisingly, social media's popularity has recently risen to the point that many use it as their principal communication route to report to the public or emergency personnel[7]. Even the CMO (Chief Marketing Officer) of many big business organizations started to respond to the question posed on social media because of its easy and global reach. According to a statistic, 40.8 percent answered on Twitter, 26.2 percent on Facebook, and 16.5 percent on LinkedIn[8]. As a result, large amounts of data are becoming a standard feature for representing civilization around the world. Several corporations have made significant investments in social media-driven decision processes in recent years, making this platform a mainstream choice for consumer data analysis and business improvement[9]. It enables businesses to engage in immediate client exposure, perhaps in the most premium way possible, resulting in higher effectiveness than conventional marketing services and technologies. The ability to evaluate, correlate, and learn from massive amounts of data is becoming increasingly vital in numerous disciplines for making a predictive decision.

The increasing volume of data generated by social media based platforms has sparked new interest in big data analytics to extract insightful meaning into the text, image, audio, video, gif, blog, etc., shared every day by billions of users[5, 10, 11]. Organizations are investing in big data analytics to research online behavior, especially on social networking sites like

Facebook, Twitter, LinkedIn, Instagram, YouTube, and blogs[1]. Many firms, through their analysts, are devoting a significant amount of time, cash, and efforts to extract valuable insights from large amounts of social data. It is necessary to employ effective procedures and analytical tools for assessing the ever-increasing data provided by numerous social media applications. The amount of study on social media has increased dramatically in recent years, and several big data analyses models have been developed to investigate more on social data analysis[4] .
The organization of this paper is as follows:
(1) Section 2 describes the background study of this research. This section investigates the concept of big data and proposes a new dynamic approach to defining big data. The "Sunflower Model" creates an opportunity to make the definition of big data up to date with new technologies. This section also explains social media, social data, social media analytic, the importance of data analysis in social media, interrelation among these elements, and many more.
(2) The research methodology is discussed in Section 3. Each of the steps in the Systematic Mapping Study (SMS) is explained in this methodology portion. The purpose of this research, previous research gap, and workflow are illustrated in this portion.
(3) The result discussion and outcomes are listed and explained in Section 4 of this study. The types, techniques, and taxonomy of big data analytics in social media are presented in this portion of the study.
(4) In Section 5, we try to evaluate this work by addressing the answer to the research questions that are set at the beginning of the SMS method.
(5) The potential challenges and limitations are discussed in Section 6.
(6) Finally, we conclude this work by giving a summary in Section 7.
perhaps in the most premium way possible, resulting in higher effectiveness than conventional marketing services and technologies. The ability to evaluate, correlate, and learn from massive amounts of data is becoming increasingly vital in numerous disciplines for making a predictive decision. The increasing volume of data generated by social media based platforms has sparked new interest in big data analytics to extract insightful meaning into the text, image, audio, video, gif, blog, etc., shared every day by billions of users[5, 10, 11].

2 Background Theory

2.1 Big data

Technological advancements produce numerous sorts of structured data, but the majority is semi-structured or unstructured data, which is mostly big data. Big data refer to large, complex data collections that necessitate sophisticated and cost-effective data administration and analysis tools to extract insights and make decisions[12] . Structured data, such as that found in spreadsheets or relational databases, account for only 5% of total data[11]. Unstructured data include online text, photographs, audio, and video, all of which lack structural organization and need special analytics/tools for data analysis[11]. A notable example of semistructured data is the Extensible Markup Language (XML), which has an informal tag-type structure for sharing data on the Web[3, 4]. Because big data are massive in quantity, too speedy, have a diverse structure, and are often sophisticated for traditional technology to acquire, preserve, maintain, and assess, it poses a significant challenge for conventional technology. The nature of large data, as well as the concerns and challenges that arise with it, hinder current data science techniques and approaches from resolving those challenges[13, 14] .
Douglas Laney is regarded as a forerunner in the fields of data warehousing and information economics (infonimics). Data strategy, big data, data analytics, infonomics, data science solutions, and so on are among his specialties. In 2001 Laney first defined big data in terms of Volume, Velocity, and Variety[15, 16]. This became the most logical and popular definition of big data (3V definition). Laney is working as a VP and chief data office in Gartner's research team[17]. Mark A. Beyer, a data scientist who specializes in data architecture, data integration, data management, and data governance, has joined Laney in this research and the two are working together on big data development[18]. As a result in 2012, Laney and Beyer increased the scope of the big data definition by adding two more V's: Veracity and Value[15, 19–21]. These two new V's are required to satisfy business objectives and goals. Without veracity and

value in data, i.e., any fake or meaningless data may lead to damage in revenue and hence degrade the decisionmaking process. Until today, 5 V's are the most widely accepted definition of big data. Recently, another group of researchers used the term "Bigs" rather than "V's" to define big data[14]. They expand the 5 V's by adding five more characteristics in big data services. The big volume, big velocity, big variety, and big veracity are grouped as fundamental features to define big data. The technological perspective of big data referred to big intelligence, big analytics, and big infrastructure. The big service, big value, and big market cover big data socioeconomically. A brief description of all bigs used in the sunflower model is written in Table 1. To reflect the combination of 5 V's and 10 Bigs in big data, we develop and propose the "Sunflower Model of Big Data". The Sunflower Model is depicted visually in Fig. 1. Each leaf of the sunflower model represents a characteristic of big data technology. We propose this as a flexible and dynamic model. This model's dynamic quality is that the leaf (new features) can be added to the sunflower to bring it up to date. However, the new feature must have a clear and logical relationship with the 10 Bigs and big data systems.

Table 1 10 Bigs including 5 V's in big data.

Bigs/V's in big data	Meaning		Remark
Big volume	This indicates the dataset's size, which is commonly measured in terabytes (TB), petabytes (PB), and other units[14]. Data volume is relative in this case, and it varies depending on a lot of things. Because storage capabilities are rising, even larger datasets will be gathered in the near future; what is described as big data now may not fulfill the barrier tomorrow[3]	1st V	Fundamental characteristics
Big velocity	Data processing speed is data velocity. The rate at which data are produced and assessed referred to as velocity[22]. It has to do with data latency and throughput[14]	2nd V	
Big variety	This refers to the wide range of data kinds, formats, and sources available. Big data can be structured, semi-structured, or unstructured, but it is mostly unstructured in practice. According to statistics, 80 percent of today's data are unstructured[14, 22] . Because social media big data are a mix of nationalities in a culturally diverse, multi-language setting, it is not structured data[14] .	3rd V	
Big veracity	It must be accurate and genuine to be deemed big data. This relates to the data's trustworthiness[22]. When working with huge data, there is confusion, imperfection, and inconsistency. However, data analytics must be used to extract meaningful insight from ambiguous, partial, and unclear big data sets[14] .	4th V	
Big intelligence	A collection of concepts, methods, and tools for managing and processing large amounts of data automatically and artificially is known as big intelligence[14]. This is a part of big computing and combined computer science, electrical engineering, mechanical engineering, data science, statistics, and so on.		

Big analytics	This is a combination of algorithms/techniques that support data management, gathering, and data analysis. Analytics uses artificial intelligence and machine learning to extract meaningful pattern which is automatic and reliable[14]. Big analytics can discover hidden patterns from unreadable raw data. Most of the time, big analytics strongly related and used big intelligence for implementation.		Technical characteristics
Big infrastructure	The architecture, tools, methods, platforms, and services that provide big data processing are referred to as big data infrastructure. The Apache Hadoop ecosystem, distributed data center, supercomputing machine, etc., are critical components of large-scale infrastructure[14] .		
Big service	A comprehensive platform capable of serving millions of individuals. Amazon web services, Google cloud services, mobile services, and social networking services are big services[14]. Often these services provide their own Application Program Interfaces (APIs) to get public access		Socio-economic characteristics
Big value	The aim of a dataset necessitates its relevance and aspect. This implies that big data bring big social value. Big data have revolutionized society in terms of socializing and perceiving, according to its high social worth[14].		
Big market	A data-driven market is required. The big market operates at a socioeconomic level[14]. This includes government, defense, education, manufacturing, business, healthcare, finance and insurance, social networking, and more.		

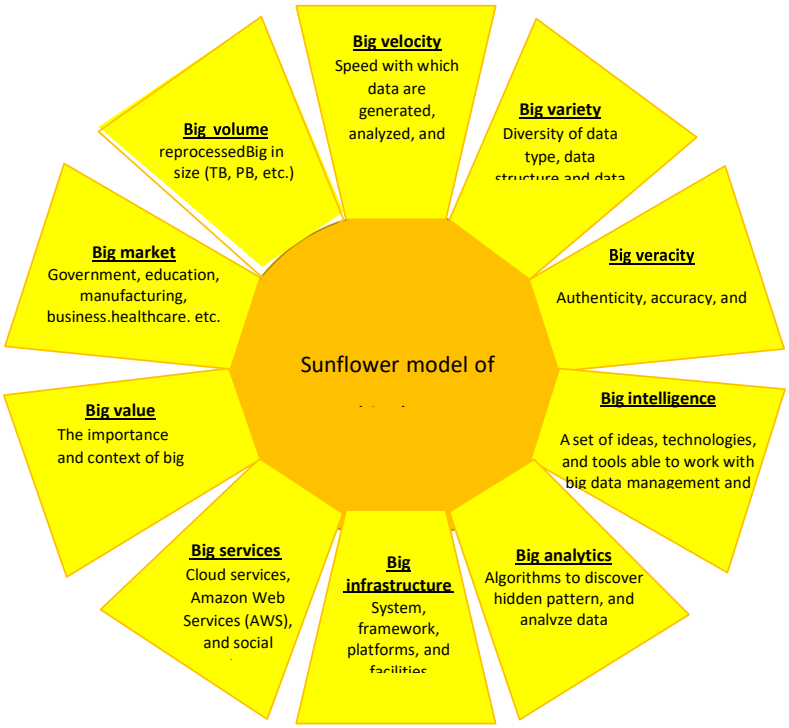

Fig. 1 Sunflower model to define big data.

It was projected that the total volume of big data is going to be 44 Zettabytes by 2020[23]. However, this was more than that in real. According to statista.com, the volume of big data reached 64.2 Zettabytes in 2020, and this will become 181 Zettabytes by 2025[24] . The fact is that data sharing on social media based platforms is continuously increasing. Every day, billions of people on social media update their statuses and post pictures and videos with their networks, revealing vital information about their interests, views, ideas, beliefs, movements, demographics, and much more[23]. The increase in data volume is also due to the pandemic's increased desire for distance learning, employment, and recreation. Furthermore, data from several other sources in the digital economy, such as smart sensors, machine logs, communications technology, geospatial data, and consumer data, is increasing rapidly[9]. Big data analytics assists scholars in evaluating structured, semi-structured, and unstructured data so that it may become useful for various companies to make important decisions. Personalization is made possible by big data analytics, helping businesses to reach out to clients in a more tailored manner based on their preferences and likes. It provides in-depth knowledge and a thorough picture of the customer, allowing organizations to personalize messages to them to increase engagement and acceptance.

2.2 Social media

Social media are internet interaction platforms that allow people to share and consume information with one another. The Programmed Logic for Automatic Teaching Operations (PLATO) system, led by the University of Illinois and marketed by Control Data Corporation, was the first social network framework for multi-communication, which was developed in 1970[25]. By permitting users to post contents that can be instantaneously edited and amended, social media began to revolutionize in 2004 with the launch of Facebook[25] . Even today, Facebook is one of the most prominent social media networks. Ellison uses three key elements to describe social media. (1) Users must be able to make their profiles public or semi-public, (2) users must be able to connect with others on the network, (3) users have access to the public behaviors of other users simultaneously[4] . Different sorts of social media include websites and tools used for social communication, forums, weblogging, social curating, social bookmarking, Wikipedia, and so on[26]. Substantial quantities of unstructured data are generated by sites like Facebook, Twitter, Instagram, LinkedIn, blogs, wikis, YouTube, and many more[4] .

This massive amount of social data can be made to speak for us. Figure 2 depicts a 60 s snapshot of social media activities from October 2021. The inspiration for Fig. 2 came from Ref. [9], and the image was taken from an online reference[27]. Figure 2 depicts the statistics of one-internet-minute data usage on some popular social media sites and the regular activities of consumers. For example, 510 thousand comments and posts were shared on Facebook, 1.3 thousand productrich pins were pinned on Pinterest, 3.47 million videos were watched on YouTube, and 210 million emails were sent in only 1 min of October 2021. This is a promising source of data that directly expresses the opinions and views of consumers on many issues. Hence, proper utilization of these data can help in critical decision making.

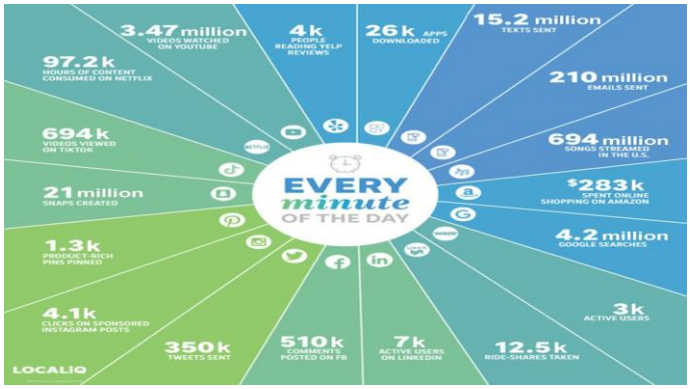

Fig. 2 Social media in 60 s from October 2021[27].

Web 2.0 technology refers to a collection of web applications that permit anyone to generate and share content online. Web 2.0 technology and high-speed internet access inspired social media towards the new concept of "Society 2.0". Online engagement by generating and sharing views among people from other societies is the motive of Society 2.0. Facebook, Twitter, LinkedIn, YouTube, Instagram, Google+, Tumblr, Flickr, emails, forums, and blogs are some of the most prominent elements of Society 2.0[9]. The social media platform is used for a variety of purposes. People use Facebook, Instagram, and Myspace for social networking; LinkedIn is popular for professional networking; Ranker website for public voting purposes; Wikipedia for knowledge sharing; YouTube and Vimeo for video content sharing; Twitter and Tumblr are used for microblogging and many more. The following Table 2 presents different types of social media identified based on their purpose. Social media are transforming our ordinary lives and creating massive amounts of data for research. Social media can help businesses make informed business decisions. Users and customers are using social media to find information and make judgments on products, education, healthcare, politicians, transportation, insurance, banks, and government services, among other things. Recent research shows that social media are increasingly being used for patient psychological health monitoring[33]. Companies are incorporating as well as using digital channels to improve organizational performance, maximize productivity, motivate staff, and collaborate with stakeholders to utilize the "Enterprise 2.0" in social networking[34]. The business has become more competitive and challenging in this digital world. Several business giants like Facebook, Google, LinkedIn, YouTube, etc., are using blessings of big data analytics on social data for predictive digital advertising, increasing inbound traffic, to enhance brand awareness and more optimizations[35]. These analytics, data, and tools boost up the marketing strategies, customer service, and global reach. As a result, large, medium, and even small businesses invest in social media data and data analytics.

Table 2 Social media types (purpose-based).

Type	Social media example
Social networking	Facebook, Myspace[28], Instagram, and VKontakte (VK)[29
Knowledge aggregation	Wikipedia[28] and Classmates[29]
Multimedia	YouTube[28], Vimeo, and Vine[29
Microblogging	Twitter[28], Thumblr, and Sina Weibo[29]
Instant messaging	WhatsApp, Messenger, Tencent QQ, WeChat, Viber, Line, and Snapchat[29]
Professional	LinkedIn, Viadeo, and Xing[29]
Forum based	Baidu Tieba (known as Postbar internationally)[29]
Communication	Skype, Hangout, and YY[29]
Social bookmarking	Pinterest[29], BibSonomy, CiteULike, Plurk[30], Delicious, Digg, and StumbeUpon[31]
Content voting	Reddit[29] and Ranker.com
Search and discovery	Google and Foursquare[29]
Friendship and dating	Tagged and Badoo[29]
Interest-based	Google+[28], beinggirl.com[31], justmommis.com, alphamom.com, and minti.com[32

2.3 Social media statistics

The abundance of consumer data makes social media a powerful resource for data analysis and research. In an unstructured format, social media content delivers a vast amount of big data. Consumer-generated social big data ensure the data's integrity and value. This makes social data more attractive to researchers, businesses, the government, and others. The majority of social media sites offer Application Program Interfaces (APIs) that allow easy and public access to large amounts of social data for research purposes. Facebook, Google, Instagram, LinkedIn, and a slew of other social media platforms provide individuals and organizations customdesigned API[31] . According to a recent Brandwatch report[2], there are 3.499 billion active social media users among the 7.7 billion global population, accounting for 45.4 percent of the overall

population and nearly 80 percent of the total of netizens. This report was published on June 13, 2019. This is due to the fact that the number of people using social media increased by 202 million from April 2018 to April 2019. Amongst 81 percent of youngsters believe that social media have a beneficial influence on personal life. Each of these users spent an average of 142 min each day on social media. This statistic shows that the retailers and businesses corporation invested 74 billion dollars on social media advertising in 2018. As a result, 91 percent of retail companies use social media sites, while 81 percent of small and midsize enterprises use social media platforms to stay competitive and maximize marketing efforts. Data generated from users of social media become so revolutionary in volume, which results in the most important source of large data. The number of engagement on various social media sites is presented in the following Table 3. This table of data is generated based on the statistics of Brandwatch[2] . Based on the number of active users, Facebook was the most popular social media platform, with 2.3 billion in a month. There were 1.9 billion active users on the video-sharing website YouTube. Similarly, WhatsApp had 1.6 billion users. The professional networking website LinkedIn had 610 million subscribers, while another major network, Twitter, had 330 million active users and so on. The percentage of active social media consumers on various platforms is presented in the following Fig. 3. Figure 3 illustrates that Facebook has the most consumers (23 percent) who are also actively participating in social communication, followed by 19 percent on YouTube, 16 percent of all are active on WhatsApp, consumers are equally active on Instagram and WeChat and so on. Individual statistics for seven of the most wellknown and rapidly growing social media platforms are presented in Table 4. Facebook, as expected, is at the top of the list. Aside from the overall number of monthly active users, Table 4 included information such as the percentage of Americans that use that platform, the average time spent by users, the percentage of business organizations who utilize social media, and more. All these statistics are collected from Ref. [2].

Table 3 Statistic of active users in social media

Social media	Number of active users (monthly)
Facebook	2.375 billion
YouTube	1.9 billion
WhatsApp	1.6 billion
Instagram	1 billion
WeChat	1 billion
LinkedIn	610 million
Weibo	600 million
Twitter	330 million
Pinterest	265 million
Snapchat	190 million
Airbnb	150 million
Google+	111 million
Flickr	90 million

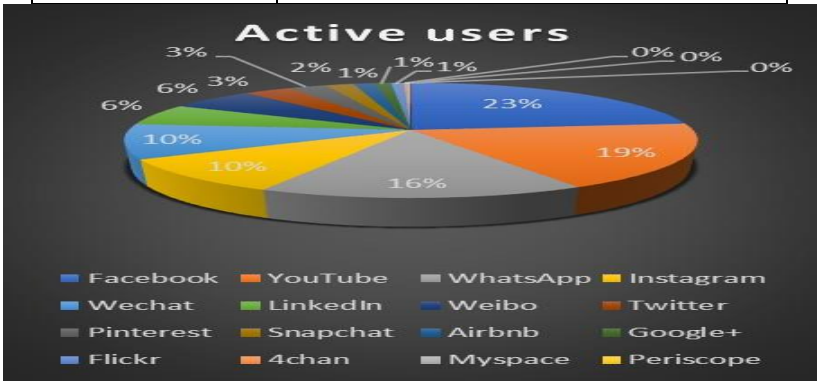

Fig. 3 Active users in social media in percent

2.4 Big social data

Every day, an enormous number of people across all social media platforms generate massive amounts of data, which can be of any type, including text, photographs, audio, video, web transactions, gifs, blogs, and other formats[5, 10, 11]. A large amount of organized, semistructured, and predominately unstructured[26] data serve as a source of big data, which is often referred to as "Big Social Data" because it originates from social media networks[34]. In short, "Big Social Data" refers to the big data gathered from and related to social media. Towards research, business, and administrative organization, the big social data are both relevant and insightful to examine and analyze to make better decisions. There are two sorts of big social data: according to some researchers, the first one is "Social Graph" and the second type is "Social Text"[37]. Another group defined what is being done in social media as "interactions" and what is being said by consumers on social platforms as "conversations"[34]. Apart from its kind, social data are messy and chaotic, containing complex social bonds like intimacy and support among consumers[28] . Traditional Relational Database Management Systems (RDBMSs) struggle to interpret social data because of the complex tree-based relationship on several data points. Traditional approaches like RDBMS do not have the ability to analyze this big social data. RDBMS is a database management system that is used to process limited portions of structured data[6]. When it comes to social data analysis, we must first choose the right big data analytics to use. Big data analytics requires the application of computer modeling that can deal with the merging of social theories with statistics and mining methodologies.

2.5 Big data analytics in social media

The systematic computing and interpretation of data using statistical methods is known as analytics. Analytics uses mathematics, statistics, and artificial intelligence to help with data analysis in difficult-tounderstand formats so that better decisions may be made. At the same time, big data analytics assists data analysis by revealing trends, patterns, and other insights from messy social data[3, 11]. In this study, the terms "big data analytics" and "social media analytics" are used interchangeably. Text mining, social graph theory, opinion mining, social influence analysis, sentiment analysis, statistical analysis, cyber risk analysis, and others are some of the diverse approaches of big data analytics in social media[28]. Furthermore, by merging, modifying, and extending ways to handle massive social data, these analytics contribute to the development and assessment of systems and informatics tools[28] .

Different firms might use the results of big data analytics to improve their production or marketing strategies to stay competitive in the digital business world. For example, social media analytics may help businesses get user input on their products, which can be used to make changes and get more value out of their brand[4, 38]. Leading companies such as Apple, Microsoft, Google, Honda, Facebook, NVidia, Amazon, Samsung, and others employ social media analytics regularly to improve their corporate strategies and customer relations practices[9]. Research, civil defense, healthcare, banking, telecommunication, public transport system, insurance, and a variety of other industries can gain benefits from social media analytics to prepare for the future and make better data-driven recommendations while remaining flexible and agile[9] . Sensitive events like elections frequently use sentiment and opinion mining in local and national elections processes[5]. The federal or state government uses social data analytics to develop a predictive decision.

Table 4 Statistics of seven well known social media sites

Social media	Individual statistics
Facebook	(1) Monthly consumers: 2.375 billion. (2) Facebook is used by 69 percent of all persons in the United States. (3) Every day, 74 percent of users visit this site. (4) On average, a Facebook consumer spends 35 min per day on the site. (5) Facebook has 60 million daily active merchant pages. (6) Facebook has 5 million active marketers.
Twitter	(1) Monthly consumers: 330 million.

	(2) Within 24 h, 500 million Tweets were sent, with 6000 Tweets being sent in less than 1 s. (3) Twitter is used by 22 percent of all persons in the United States. (4) Companies in the United States have begun to invest in digital marketing, with 65.8 percent of those with 100 or more employees using Twitter. (5) According to a recent study, 77 percent of Twitter consumers feel positive about a brand on receiving comments on their Tweets.
YouTube	(1) Monthly consumers: 1.9 billion. (2) In less than 1 min, 300 h of content were posted to YouTube. (3) Each person consumes 40 min of YouTube videos per day on average. (4) YouTube is consumed by 90 percent of Americans between the ages of 18 and 24. (5) YouTube is used by 9 percent of small businesses in the United States
Pinterest	(1) Monthly consumers: 265 million. (2) The platform is used by 28 percent of all Americans. (3) The male audience increased by 41 percent in 2014, and the number of hours spent on Pinterest tripled to over 75 min by each visitor
Instagram	(1) Monthly consumers: 1 billion. (2) Every day, about 95 million photographs are shared. (3) As of now, over 40 billion pictures have been posted. (4) 37 percent of all persons in the United States use Instagram. (5) Each Instagram user occupies 15 min a day on the platform.
LinkedIn	(1) Monthly consumers: 610 million. (2) The platform is used by 27 percent of adults in the United States. (3) LinkedIn profiles have been created by over 3 million businesses. (4) LinkedIn is used by 17 percent of smaller businesses in the United States

3 Research Methodology

This research is carried out using the Systematic Mapping Study (SMS), a scholarly and well-known methodology in the field of scientific surveys. Another name of the Systematic Mapping Study is Systematic Review (SR). Sequential activities by following a series of independent tasks lead to the ultimate goal in this system. The series of tasks from the SMS is mostly used to collect and scrutinize scientific articles from a certain topic area to answer some predetermined questions[39]. The strategy behind this is to find and assess all applicable articles to address specific problems[24]. We use the guidelines advocated by Kitchenham and Charters[40] and Petersen et al.[41] to implement the SMS method. Although to serve the research purpose, we slightly modify the tasks of SMS like Refs. [15, 19]. The following six tasks are followed sequentially to conduct this research: (1) research goal; (2) research questions; (3) searching strategy; (4) selection criteria; (5) selection of studies; and (6) result analysis.

3.1 Research goal

There is no doubt that social media have become good source of big data and data analysis research. Texts, images, audios, and videos shared by social media users generate a huge amount of structured, semistructured, and unstructured raw data daily. These data generated by the active participation of end-users attract business organizations, researchers, educators, and even governments to the social data analysis. Adjustability and variety in big social data require data analytics, machine learning, and data science in the domain of social media data analysis. There is much analytics for big data analysis, but not all of them can be used to analyze data on social media. With great concern, there is a paucity of research that covers all aspects of big data analytics and includes a variety of implementation strategies and algorithms. To fill this need, we have set out this research to (1) identify big data analytics and (2) their associative

algorithms that go with analytics for social media data analysis. Besides, we want to investigate the supporting data type (e.g., structured and unstructured) of each of these big data analytics. So that we can present a clear data view to other researchers in this domain. Another goal of this research is to show how social media are good source of big data based on real-life statistics. We will become acquainted with big social data, society 2.0, and social data analytics as a result of this research.

3.2 Research questions

(RQ) Based on the purpose of this research, we set up the following RQ. This assists in the formulation and acceptance of this scientific study.

RQ1: What are the popular analytics used in the social media based platform to analyze data?

RQ2: Which techniques/machine learning algorithms are used to implement these analytics in social media?

RQ3: What are the supporting data types of these analytics?

RQ4: Which one is the most popular big data analytics in social media data analysis?

3.3 Searching strategy

We create a strategy for searching scientific resources based on key words to find solutions to the above study questions from various relevant resources. We plan very brief and specific key words for searching. Our searching key words are

"Big Data" or "Social Data"

And

"Analytics" or "Data Analytics"

and

"Social Media" or "Social Network" or "Social Sites".

We choose to use these key words only on the article title and sometimes on the article abstract. The ACM digital library[42], IEEE Xplore digital library[43], and the ScienceDirect digital library[44] were chosen to search scientific papers by using these key words.

3.4 Selection criteria

We devise a set of criteria to include the most important scientific papers while excluding those that are unrelated or minorly relevant. The Inclusion Criteria (IC) are as follows.

IC 1: This specific study relates to big data, data analytics, and social media.

IC 2: This specific study is a published scientific paper.

IC 3: Full text is available.

IC 4: This article is from the computer science domain.

IC 5: This is a review/research article. In the same way, the Exclusion Criteria (EC) are as follows.

EC 1: This specific study is a summary of a conference/workshop.

EC 2: This specific study is a course or a book chapter.

EC 3: The full text of this study is not available.

EC 4: This specific study represents big data analytics but does not relate to social media.

EC 5: This is video content. EC 6: Papers written in a language other than English.

3.5 Selection of studies

For this study, we select articles from three popular research databases. These are ACM, IEEE, and Science Direct digital library. We use the advance search and filter option provided by each of these databases. Only ACM Transactions On Knowledge Discovery From Data section is evaluated from the ACM digital library to eliminate overlaps in articles. The number of articles from each of these three databases is depicted briefly in Table 5.

We receive several duplicates among these articles. To select best-fitted articles, we remove duplicates and then apply the exclusion and inclusion criteria to minimize the number of articles for this study. A total of 85 articles are chosen at the beginning. After a thorough reading of the title and abstract, the primary list is minimized into 41 articles. After a deep investigation on the result part of each of these 41 articles, a total of 20 papers have been chosen as found very relevant to this research. This entire process workflow is depicted in Fig. 4.

Table 5 Number of articles per database

Source database	Number of papers

ACM digital library (ACM)	502
IEEE Xplore (IEEE)	375
ScienceDirect-Elsevier (SD)	625

4 Result

4.1 Social media analytics

The outcomes from the systematic review of this research endeavor are described in this part. The primary goal of this research is to identify and collect notable big data analytics used in social networking platforms. We get 10 mostly used social media analytics so far. The details list is presented in Table 6. The serial numbers of source papers are mentioned in the leftmost column. The second column lists the titles of the final 20 articles in this study, along with their reference numbers. The authors of those publications, as well as the year of publication, are included in the third column for clarification. Finally, the right-hand column displays the name of Big Data Analytics (BDA), as discovered in those twenty articles. Different analytics are used for different purposes in the domain of social media. For example, text analytics for text analysis, video for video data analysis, and image data are analyzed by using image data analytics. To date, "Text Analytic" has been the most widely used and chosen analytic method for large-scale social data analysis.

4.2 Taxonomy of analytics in social media

For data analysis, we discovered that ten big data analytics are mostly discussed in the social media sector. These social data analytics are divided into three groups. These are (1) based on data types, (2) based on purpose, and (3) based on the nature of the task. The taxonomy is shown in Fig. 5. There are four social media analytics based on the data type. These are primitive data types like text, image, audio, and video. (1) Text analytics works with string/text data. For example, review on a consumer product, comment on a topic, views on an issue, and other text data from social media. (2) Image analytics supports images, pictures, scenario, or photographs of any object. Social media users enormously share a picture of a business product, a beautiful moment of a trip, photographs of events, or a social gathering. (3) Audio analytics uses machine learning to extract meaningful information from audio, speech, or music. Several kinds of research go on to convert speech into text, analyze audio of social media users to extract insights, and others. (4) Video analytics shows the recent advancement of technology in social data analysis. Making video data talk for us is a new era in digital communication and data assessment.

Based on the purpose of data analysis, there are another four types of social data analytics. (1) Predictive analytics uses a machine-learning algorithm to develop a forecasting model. This model gives data prediction based on historical data analysis. (2) Descriptive analytics identifies flaws by analyzing data from the present or past. This analysis assists in monitoring events and generates results in the form of a report. (3) Prescriptive analytics examines several situations and offers the most optimal solution. This emphasizes conditions and critically chooses the best outcome based on the historical condition-result relationship. (4) Diagnostic analytics works continuously to develop better results. Data mining and data correlation assist in each round of diagnostic improvement in the social data analysis process.

To do other specific tasks in social networking platforms, there is two more big data analytics. (1) Visual analytics expands the concept of video analytics. This works with video, image, animation, gif, and other forms of visual data. Social Set Visualizer (SoSeVi) is a good example of visual analytics[37, 49]. (2) Web analytics is some analytic tools provided for free and public use. The data from WWW that are automatically generated or indirectly connected with users like metadata, log file analyzer, transaction on web, bookmarks data, etc. are an example of data used in web analytics. Web analytics works with other social data, too.

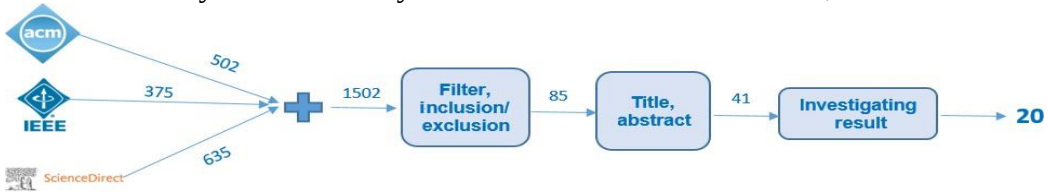

Fig. 4 Selection process for this survey.

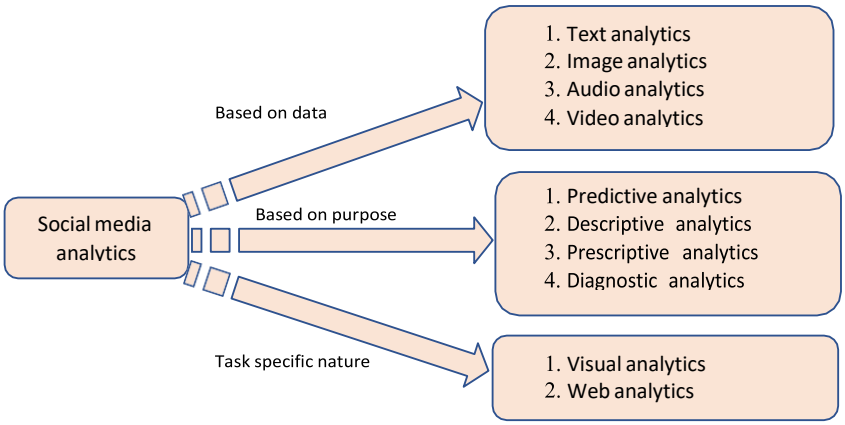

Fig. 5 Taxonomy in social data analytics.

4.3 Machine learning techniques in social media analytics

In the context of social media, different algorithms are used in association with distinct types of data analytics. Table 7 shows all of the techniques employed with each of the 10 Big Data Analytics (BDA) classes revealed in this study. The serial number of the BDA and the name of the BDA are stated in the leftmost column of Table 7. The middle column lists associated statistical or machine learning methods/techniques with the relevant social data analytics. The scope of machine learning algorithms definition is loosely considered rather extended for the sake of the articles found in this study. Machine learning algorithms, to broaden the scope of work, include not only mostly used algorithms but also a technique, or an approach, or a procedure that may employ an algorithm in the background to evaluate any kind of social media data. For example, there are some similarities among sentiment analysis, sentiment classification, and social network analysis but they all differ by approach, purpose, and procedural way behind them. Sentiment analysis can be done by both supervised and unsupervised learning methods, while the sentiment classification must follow a supervised learning method, on the other hand, social network analysis follows a graph theory to analyze social data[51–54]. The goal and data analysis techniques are different in each of these three methods. Similarly, Google Analytics is a web analytics technique to track and report website traffic[55, 56]. Many business organizations frequently use google analytics for online business and marketing purposes. AWStats, Amung.us, and WebSTAT are other similar tools where machine learning algorithms are working from behind. Most of the researchers use these tools and techniques as the brand name rather than the behind algorithms or combination of algorithms. To increase the clarification, we listed the name of techniques and the machine learning algorithms in a broad sense. Popular machine learning algorithms are included as well like Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Naive Bayesian classifiers (NB), Random Forests (RF), Decision Tree (DT), and many more.

Social media analytics support structured, semistructured, and unstructured data types. The rightmost column of the following Table 7 presents which BDA support whatever data type for social data analysis. Text analytics supports both structured and unstructured formats of data. Derived numbers from social text data are in structured data format, while text data are unstructured. Image analytics, audio analytics, and video analytics mostly work with complex, unstructured, and messy data. In this study, we find that predictive analytic and descriptive analytic support both structured and unstructured data types while diagnostic and prescriptive mostly work with only unstructured data. Visual analytics always works with the unstructured data type. Web analytics can work with structured, semi-structured, and unstructured data. These strategies are crucial for enhancing decision-making by analyzing a large amount of potential social data. As a result, these methodologies represent a useful subset of the big data analytics technologies accessible to the researchers.

5 Evaluation of Study

We specified four research questions to evaluate this research project. Gradually, we address these questions and try to achieve the goal we established at the start of the project.

Table 7 Techniques used in each of social media big data analytics.

BDA type	Technique or algorithm	Working data type
BDA 1: Text analytics (text mining) / Text classification in Refs. [1, 3, 4, 6–9, 26, 28, 31, 33–35, 45, 47–50]	Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Gibbs Sampling Approach, Latent Dirichlet Allocation Algorithm, Random Forests (RF), Decision Tree (DT), Information Extraction (Entity Recognition and Relation Extraction), Sentiment Analysis/Opinion Mining (Document Level, Sentence Level, Aspect Based, Location (Country) Based, Timestamp Based, and Followers Count Based), Lexical Resource Approach, Probabilistic Neural Network, Unstructured Data Normalizer (UDN), Text Summarization (Extractive Approach and Abstractive Approach), Social Influence Analysis, Natural Language Processing (Information Retrieval based Approach, Knowledge based Approach, and Hybrid Approach), Social Data Analytics Tool (SODATO), Support Vector Machine (SVM), Nave Bayesian classifiers (NB), Logistic Regression (LR), Multinomial Logistic Regression, Restricted Boltzmann Machine, Message Content Analysis, Non-parametric ANOVA Analysis, Cluster Analysis, Cluster Dendrogram Analysis, Histogram Analysis, Word Cloud and Commonality analysis, Pyramid Analysis, Cyber Risk Analysis, Social Network Analysis, Statistical Analysis (Markov chain Monte Carlo methods, regression models, and factor analysis), Trend Analysis, Extended Log File Analyzer (cross correlation, self-updating system, customize the configuration, Near Real Time Extensions (NRTE)), Social Media Product Improvement Framework - SM-PIF (Contextual Information Retrieval (Feature Based Ontology (FBO) and Extraction and Storage (ES)), Feature Improvement Recommendation (Product Recommendation Service (PRS)), Artificial Neural Networks (ANN), Swarm Intelligence, Evolutionary Computation, Deep Learning, Formal Model, and Fuzzy Logic	Structured and unstructured
BDA 2: Image analytics (image classification)[7, 31]	Convolutional Neural Networks (CNN), Support Vector Machine (SVM), Linear SVM, Statistical Analysis of tag data, demographic data, download frequency, etc.	Unstructured
BDA 3: Audio analytics (speech analytics)[3]	Transcript-based Approach (large-vocabulary continuous speech recognition (LVCSR)) and Phonetic-based Approach	Unstructured
BDA 4: Video analytics (Video Content Analysis (VCA))[3, 31, 46]	CTV metadata analytic, Modified CCTV VMS (Video Management System), Serverbased Approach and Edge-based Approach, and Statistical Analysis by number of users, response rate, subject, and location	Unstructured

RQ1: What are the popular analytics used in the social media based platform to analyze data?

The answer to this question is given in Section 4.1. One of the main goals of this research is to determine which big data analytics are most commonly utilized to evaluate social data. Table 6 in Section 4.1 lists the social data analytics mentioned in each of the articles chosen for this study. In this study, ten of the most prominent big data analytics in the realm of social media data analysis and decision-making are selected. These ten are, (1) text analytics, (2) image analytics, (3) audio analytics, (4) video analytics, (5) predictive analytics, (6) descriptive analytics, (7) diagnostic analytics, (8) prescriptive analytics, (9) web analytics, and (10) visual analytics.

RQ2: Which techniques/machine learning algorithms are used to implement these analytics in social media? Big data on social media are complex in structure to analyze by a traditional simple model. Big data analytics for social data analysis is aided by many statistical or machine learning methods. The outcome of this study on associated techniques/methods with each data analytics is presented in Table 7 of Section 4.3. Table 7 lists algorithms/methods used to implement social data analytics and data analysis that address the answer to RQ2.

RQ3: What are the supporting data types of these analytics? In Section 4.3, the third column of Table 7 lists which social media analytics support what types of data. Structured, semi-structured, and unstructured data are the three forms of data we are familiar with. To keep things simple, we simply consider and list structured and unstructured data in this study. The last paragraph of Section 4.3 apparently presents a short description of which big data analytics mostly works with what type of data in terms of the structured and unstructured format. To make this easier to understand let us discuss the data type in terms of text, image, audio, video, etc. Text analytics always works with text or string type of data[1, 3, 4, 6–9, 26, 28, 31, 33–35, 45, 47–50]. Sentiment analysis is the widely used method for text data analysis till now. Image analytics supports the image, picture, or snap of an object [7, 31]. Object detection or image classification is one of the most frequent tasks of image analytics. Audio analytics supports any kind of audio or speech from social media while video analytics works with video data in any format[3, 31, 46]. These two are emerging fields of data analysis in social media. Researchers are always interested to work on predictions based on data beyond any format. This is found that predictive analytics frequently works with text, image, audio, video, and other formats of data[3, 4, 33, 34, 37]. In this study, we find descriptive analytics mostly used to analyse. numbers, text, and images[4, 34, 37]. Similarly, diagnostic and prescriptive analytics is mostly used to handle text data[4, 34, 37]. Web pages are the giant source of semistructured data. Web analytics is several tools supported by a blend of multiple techniques to analyze web traffic and web data[1, 9]. Visual analytics is a special type that does not work with video data but supports image and animation types of data for analysis[34, 37, 47, 49] .

RQ4: Which one is the most popular big data analytics in social media data analysis? “Text Analytics” remains the most popular and widely utilized data analytics in the social media sphere. Text analytics as a computational tool in social data analysis is mentioned in 90 percent of papers in this study. The second prominent social data analytics is “Predictive Analytics”, 25 percent of the total articles of this survey describe predictive analytics as a tool for social data analysis. Five percent of total work in social data analysis is being done by “Diagnostic Analytics” and “Audio Analytics”. The usage of different social data analytics is presented in the following Fig. 6.

6 Challenges and Limitations

Many disciplines and sectors have advanced as a result of the widespread use of social media data and big data analytics. There are numerous hurdles and limitations to working in this field.

1.With the increasing abundance of social media data, files are now being distributed over multiple physical sites. Public access is becoming difficult and technical skill is needed to access these data.

2.The maintenance of large social datasets is challenging and expensive.

3.Consumes continuously sharing status updates, photos, videos, etc., are not always useful for analysis. Data cleaning and filtering are required to extract necessary data from this complex dataset that is costly and time-consuming.

4.Cyber-attacks have a severe impact on social data during sensitive events such as elections, which could result in a faulty conclusion.

5.There is a great chance of getting unreliable and incomplete data. Like noise, misspelling, foreign data in English, etc.

6. Integrating and combining social data from many platforms is a difficult task.

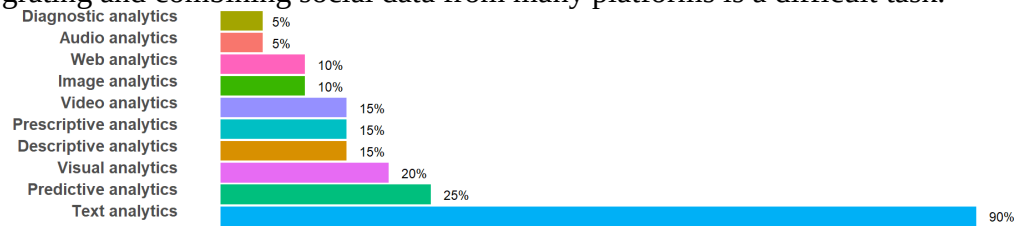

Fig. 6 Utilization of social data analytics.

7 Conclusion

Big data, along with advances in computing tools, have evolved as a significant data analytics for understanding human behavior by analyzing data from social media. All types of organizations, from industry to government, can be benefitted using social data, data science, and social data analytics. This study fills in the research gap by identifying the ten most widely accepted and used big data analytics for analyzing social data and making decisions. Considering the overlap among the approaches of social media analytics, we design a taxonomy of big data analytics in the social media domain. We create three main categories and then assign these ten analytics to each one depending on the purpose, nature of usage, and working area. Data analysis in social media is aided by machine learning techniques. Each of these social data analytics has a long list of machine learning or statistical methodologies associated with it. We present social data analytics along with the methodologies. Until now, the most widely utilized analytics for social data analysis has been “Text Analytics”. In addition, researchers are advancing their ability to extract useful information from an image, audio, and visual material in social data. The social media platforms provide data continuously to evaluate because it allows people to offer their perspectives on the most current event, product, tools, talents, and other topics. We must take advantage of the benefits of this massive amount of data.

This research looks at big data analytics in social media in a broad, generic way. A specific field of interest, for example, business analytics in social media, geospatial/location based analytics, social media data analysis for political science research, etc. can be explored to serve the same purpose. We will continue our investigation by focusing on a small number of social media platforms, such as Facebook, Twitter, and Snapchat. Any of the ten big data analytics described above can be explored further. We do not get enough time and resources to investigate deep on the list of machine learning algorithms on each big data analytics. We aim to keep working on this project to find a shortlist of acceptable algorithms for each of these ten big data analytics categories. We also want to figure out and decide a few common attributes/characteristics of big data analytics by which we can tune up one analytics and perform comparative performance analysis.

References

[1] V. Dhawan and N. Zanini, Big data and social media analytics, Res. Matters A Cambridge Assess. Publ., no. 18, pp. 36–41, 2014.

[2] K. Smith, 126 amazing social media statistics and facts, https://www.brandwatch.com/blog/amazing-social media-statistics-and-facts/, 2019.

[3] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015.

[4] N. A. Ghani, S. Hamid, I. A. Targio Hashem, and E. Ahmed, Social media big data analytics: A survey, Comput. Human Behav., vol. 101, pp. 417–428, 2019.

[5] P. V. Paul, K. Monica, and M. Trishanka, A survey on big data analytics using social media data, in Proc. 2017 Innov. Power Adv. Comput. Technol. (i-PACT), Vellore, India, 2017, pp. 1–4.

[6] F. Shaikh, F. Rangrez, A. Khan, and U. Shaikh, Social media analytics based on big data, in Proc. 2017 Int. Conf. Intell. Comput. Control. (I2C2), Coimbatore, India, 2017, pp. 1–6.

[7] V. Nunavath and M. Goodwin, The role of artificial intelligence in social media big data analytics for disaster management–initial results of a systematic literature review, in Proc. 2018 5th Int. Conf. Inf. Commun. Technol. Disaster Manag. (ICT-DM), Sendai, Japan, 2018, pp. 1–4.
[8] F. Piccialli and J. E. Jung, Understanding customer experience diffusion on social networking services by big data analytics, Mob. Networks Appl., vol. 22, no. 4, pp. 605–612, 2017.
[9] P. Ducange, R. Pecori, and P. Mezzina, A glimpse on big data analytics in the framework of marketing strategies, Soft Comput., vol. 22, no. 1, pp. 325–342, 2018.
[10] P. Grover and A. K. Kar, Big data analytics: A review on theoretical contributions and tools used in literature, Glob. J. Flex. Syst. Manag., vol. 18, no. 3, pp. 203–229, 2017.
[11] J. Amudhavel, V. Padmapriya, V. Gowri, K. Lakshmipriya, K. P. Kumar, and B. Thiyagarajan, Perspectives, motivations, and implications of big data analytics, in Proc. 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET), Unnao, India, 2015, pp. 1–5.
[12] M. Gupta and J. F. George, Toward the development of a big data analytics capability, Inf. Manag., vol. 53, no. 8, pp. 1049–1064, 2016.
[13] L. Cao, Data science: Challenges and directions, Communication of the ACM, vol. 60, no. 8, pp. 59–68, 2017.
[14] Z. Sun, K. Strang, and R. Li, Big data with ten big characteristics, doi: 10.13140/RG.2.2.21798.98886.