Global Journal of Information Technology: Emerging Technologies

Volume 10, Issue 2, (2020) 128-145

www.gjit.eu

Conceptual usage model of big data generated by social media

Jolanta Sabaitytė*, Department of Defence Technologies, General Jonas Zemaitis Military Academy of Lithuania, silo str. 5a, 10322 Vilnius, Lietuva

Roberta Karpovičiūtė, Faculty of Business Management, Department of Business Technologies and Entrepreneurship, Vilnius Gediminas Technical University, Sauletekio Ave. 11, 10223 Vilnius, Latvia

Suggested Citation:

Sabaitytė, J. & Karpoviciutė, R. (2020). Conceptual usage model of big data generated by social media. *Global Journal of Information Technology: Emerging Technologies. 10*(2), 128–145. https://doi.org/10.18844/gjit.v10i2.4764

Received from June, 13, 2020; revised from August 18, 2020; accepted from; October19 , 2020.

Selection and peer review under responsibility of Assoc. Dr. Devkan Kaleci, Inonu University, Turkey.

Abstract

The digital revolution and the communication platforms provided by the Web 2.0 virtual space era, such as social media, social networks, other tools and channels, create new opportunities for better marketing decisions based on user-generated data analysis. Every day customers of social media and other virtual tools are creating huge amounts of their actions-caused data, and businesses lack management tools supporting this process, which could create knowledge in the areas of deeper cognitive customer profiles and preferences. The growing number of social media users indicates the popularity of these communication tools among the information society, but science today lacks a deeper knowledge of social media-generated data and other algorithms for this kind of data usage. Therefore, the purpose of the article can be defined as the development of a conceptual model of big data generated by social media usage in business. The formation of the conceptual model is based on the analysis of big data assumptions and application possibilities, social media classification peculiarities and different channel specifics, identification of big data analysis methods and analysis of big data applications generated by social media. The conceptual model creates preconditions for deeper knowledge of user-generated big data in nowadays' widely used communication platforms, as well as creation of the decision support tool for marketing specialists in order to use big data from social media in deeper cognitive customer profiles and preferences. The methods employed in this research are literature and other references analysis, synthesis and logical analysis of information, comparison of information, systemisation and visualisation.

Keywords: Big data, data mining, social media, social networks, internet marketing.

* ADDRESS FOR CORRESPONDENCE: **Jolanta Sabaityte**, Department of Energy, University of Kirkuk, Kirkuk 36001, Iraq.

E-mail address: jolanta.sabaityte@lka.lt

1. Introduction

Technological advances and increased information volumes have facilitated the accumulation and use of big data. The speed and growth of data generation are driven by the proliferation of mobile devices and constantly growing internet speed. The number of internet users has already exceeded four billion worldwide, which means that more than 55% of all humanity's population are using the internet [30]. Rapidly rising social networks and social media have a great impact on big data projects as the popularity of these technologies are growing and the data, generated by social media, are rich in information. According to Statista [52], every minute internet users view more than four million YouTube videos, make more than three million searches on the Google search engine, share over two million photos on the social network and do a lot of other interactions that generate huge amounts of data. Moreover, in 2017, it was predicted that by 2021 the number of social media users will exceed three billion; however, according to the We Are Social report [57], this number was exceeded in 2018, which shows that it is obvious that the amount of data generated by social media will grow faster and faster [51]. Analysing such data provides valuable insights as it can reflect consumer behavioural trends and other public information that enables more accurate analysis [7]. A company that knows its customer has a great competitive advantage because it can make better decisions and develop products more efficiently based on the collected data. For such organisations, it is easier to communicate with existing or potential users, understand their needs and predict future behaviour. Moreover, organisations that are focusing on audience analysis based on data mining can spend their advertising budgets more efficiently and achieve better results than their competitors. Data from social networks and search engines are not only growing in quantity and variety, but also expanding the boundaries and capabilities of big data as it is being connected with new platforms and channels. However, there is a lack of tools for understanding the value created by big data and for creating the conditions for using it. The problematic issue of this work is the potential of big data generated by social media in marketing. The aim of the thesis is to form a conceptual framework for the use of big data generated by social media. The following tasks have been set to achieve the goal:

- To analyse the concept of big data and the development of the concept;
- To analyse the definition of social media data and the classification paradigm;
- To analyse the possibilities of using data generated by social media;
- To create a conceptual model for the use of big data generated by social media.

Literature and other references analysis, synthesis and logical analysis of information, comparison of information, systemisation and visualisation methods are employed in this research.

2. Big data characteristics and evolution of the concept

Increasing automation, internet speed and the proliferation of smart devices all contribute to the continuous growth of data volumes. The amounts of data have become so large that simple databases are not capable to manage them, as well as traditional methods and programmes are no longer proper for storing and analysing such data. To define these large amounts of data, the term 'big data' was formulated.

It is said that the term 'big data' was first mentioned in 1998 during the presentation by John Mashey of Silicon Graphics. Later, this term was also used in academic literature [59]. Currently, the term 'big data' is widespread; however, various definitions of big data can be found in the literature, and different authors formulate this concept based on various aspects, for example:

- Oussous et al. [42] define big data as large, fast-growing amounts of data that are composed of heterogeneous formats: structured, unstructured or semi-structured data that require powerful technologies and advanced algorithms, as traditional data-retrieval methods for such data are no longer effective.

- Alharthi et al. [6] define big data as large amounts of large-scale data created by human activity, which is difficult to manage using traditional analytical tools.
- Politaitė and Sabaitytė [44] claim that big data are large data arrays for which conventional data processing tools are not proper.
- As defined by Yin and Kaynak [61], big data are data sets whose size exceeds the ability of the conventional database software to accumulate, store, manage and analyse.
- Sharma [48] also emphasises that these are data sets that are so large and complicated that they become difficult to process using traditional data processing applications and tools.

It can be noted that the authors emphasise the size, growth and speed of this data, as well as the fact that it is difficult to process using traditional data analytics tools. Some authors define big data based on the three main characteristics attributed to the 3Vs model: volume, variety and velocity [6], [37], [46], [53].

- Volume defines the size of the data that is stored. Big data can be provided in terabytes, petabytes, zetabytes or exabytes. However, the definition of big data size is relative and may vary depending on the data type or sector of use [23]. For this reason, there is no precise definition of what amount of data is required for data to be considered as big data. Also, because of the growing amount of data and advanced technologies, the perception of big data is changing, that is why data which is considered as big data today can no longer be considered as big data later. It is estimated that in 2012 about 2.5 exabytes of data were generated each day and these figures doubled over 40 months [39]. According to the International Data Corporation, in 2013, this number exceeded four zetabytes [45]. Moreover, it was predicted that in 2020 the data volume will reach 40 zetabytes [36].
- Variety defines the heterogeneity of collected data [53]. This feature highlights the fact that big data can be generated from multiple sources and formats, including structured and unstructured data [46]. However, majority of the data (95%) are unstructured, including data generated by social networks, sensors and smart devices. Although companies have previously collected large amounts of data from internal and external sources, the use of big data has changed as new technologies are used to process and analyse data [23]. In other words, technological advancements enable companies to collect, process and analyse big data more efficiently.
- Velocity refers to the speed of data generation (how often the software generates new data), the speed at which it should be analysed and processed and how fast data is transmitted to the analytical data set after the data is generated [23], [46]. The growing use of internet-enabled devices and social media is leading to the speed of data collection and the need for real-time analysis and evidence-based planning. Data generated by smart devices and mobile applications generate high-speed information flows that can be used to analyse data in real time, thus creating customised offers for existing and potential customers [23].

In newer literature, the authors have broadened the concept of big data with the 5Vs model, which is complemented by the 3Vs model with two new components – veracity and value [48], [61]:

- Veracity is important as the majority of data coming from the internet or various sensors is incorrect [45]. Data collection should include requirements that define data accuracy and correctness. Therefore, it is important that erroneous data would be removed before the analysis processes begin. The company should define what data are needed and for what purpose, thus reducing the amount of false or unnecessary data [61]. Data flows are growing rapidly, and large volumes are more difficult to process, requiring more resources and leading to misconceptions (the greater the amount of data, the bigger the possibility of false information).

- Analysing the collected data provides valuable insights on how companies can adjust their strategy, improve their production, get to know their client, create personalised messages and gain a competitive edge. This leads to another characteristic that is value. It is important to understand that the data itself is not valuable, the value is obtained by analysing it [45]. Cavanillas et al. [15] have provided the data value creation chain which contains the processes of data collecting, analysis, mentoring, storage and use (Figure 1). As a data analyst can track business trends in a variety of queries and filter the data, the value can be described as a key feature of big data [27], [49] and the final product.

Figure 1. The chain of big data value creation (created by the authors, based on [15])

Analysing the concept of big data, 6Vs [25] and 7Vs [23] models with validity and visibility characteristics were offered:

- Validity describes the logic of the data [35]. At first sight, this characteristic may be associated with veracity. However, when speaking about validity it is more important to know whether the data has meaning assigned to it or whether it corresponds to the facts. The need for justification can be linked to the protection of personal data, as it is important to ensure that the source of the data is transparent and appropriate for storing such data.
- Visibility is also important to ensure that the data are reliable and accessible. This feature allows the user to see and analyse the data with their access [23].

It can be seen that the V models are evolving – they are complemented by new elements based on the big data properties. Considering that the V models are based on the characteristics of the big data, it would be appropriate to treat them as models of big data characteristics (Figure 2).

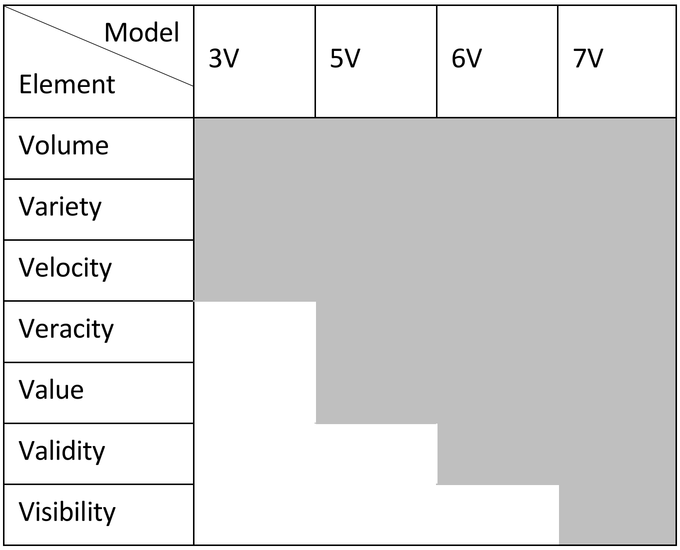

Figure 2. The elements of big data characteristics models (created by the authors, based on [6], [23], [35], [46]–[48], [53], [61])

By analysing the scheme, it can be observed that big data characteristics models include elements of volume, variety, velocity, value, veracity, validity and visibility. In the figure, the grey colour is a representation of elements specific to a particular data model. With the evolution of big data models, the old elements remain as important, but they are complemented by new elements that define big data not through the prism of their properties but are more focused on data quality.

To sum up, it can be said that big data is a large variety of high-volume, fast-growing data that requires advanced storage and processing technologies and that can provide valuable insights. However, in order to derive value from this data, it is important that data would meet conditions of fairness, reasonableness and visibility.

3. Classification of big data

To better understand the characteristics and diversity of big data, it is important to define the possible types of big data. Big data can be categorised according to their specific characteristics, such as source or data format.

Some of the authors [13], [21], [22], [23], [6] identify possible data sources as follows:

- *Social Media*. It is a source of information that allows users to share or exchange information and ideas in virtual communities, such as blogs, microblogs or social networks [27]. Social media generates such data as social networking links [8], [16], photos, comments, headlines [56], video records, reactions (such as 'likes' on Facebook or sharing other users' content).
- *Smart devices*. Data from smart devices (phones, computers and other digital devices) are attributed to machine-generated data – data that are automatically generated from devices or software such as computers, medical devices or any other devices, without human intervention [27].
- *Internet of Things devices*. Numerous smart devices connected to the internet, such as smartphones, tablets and digital cameras, interact with each other to generate huge amounts of data [27]. The Internet of Things technology describes the relationship between physical things and people in the world, combining technologies and platforms.
- *Sensors* are used to transmit signals that can encrypt physical data (e.g., photoresistors transmit solar-related data).
- *Transaction data* (e.g., financial transaction data, medical records, e-commerce website transactions, such as orders and payments).

Most authors [23], [27], [46], [61] introduce the following types according to the big data format:

- Structured data includes data related to databases that have a structure [53]. Structured data can be numbers, words, dates and data tables. Such data are easy to input, store and analyse.
- Unstructured data include data collected using modern technology, such as photos, graphics, audio files, comments, email, blog posts and other types of files. Unstructured data represent 95% of all data [23].
- Semi-structured data include data that do not conform to the strict standards covering fully structured and unstructured data [53]. XML files are assigned to semi-structured files.

Thus, in the literature, most of the data are classified based on the source or data structure. Structured and unstructured data require different processing methods. It is also important to emphasise the fact that unstructured data are more difficult to analyse, it requires more advanced technologies, but its processing can provide more valuable insights, allowing us to understand the consumer's behaviour and interests better. Using data mining technologies and algorithms from unstructured data, structured data can be extracted [23].

Moreover, social media data can be divided into two more groups [23]:

- User-generated content, which is data that include user-published content, such as photos, videos, reviews, comments etc.;
- Data based on user relationships and interactions.

These aspects determine which perspectives are used to analyse the available data – the user-oriented content analysis, which focuses on content, e.g., analysing photos, videos, reviews and comments, and structural analysis, where data based on user relationships are analysed.

Thus, when classifying and distinguishing big data, it can be observed that different data groups require different analytical methods. Also, when planning big data projects, organisations can accurately predict what data should be collected and how it should be planned for further processing and analysis operations.

Until now, social media has been seen as a communication channel, but nowadays more and more attention is being paid to analysing social media as a data source and identifying opportunities for their use. Therefore, the following section will analyse the potential of using big data from this source and data mining methods.

4. The concept of social media classification

With the rapid growth of social media popularity, the data volumes generated by social media and usage opportunities are growing as well. From 2010 to 2018, the number of social media users has increased more than three times – in 2010 there were 0.97 billion social media users worldwide [51] and in 2018 this number exceeded 3.19 billion [57]. Social media includes not only sources of information, but also collects and stores huge amounts of information related to consumer behaviour online [4]. Based on the comments, reactions to posts, news that users follow or by who they are being followed, data analysts can say a lot about users and their behaviour. With big data coming from social media, marketing professionals can gain useful insights and implement data-based marketing strategies. The growing importance of social media data is also supported by the growing number of searches in this area. Based on Google Trends data, since 2013 the number of searches related to big social media data has doubled (Figure 3).

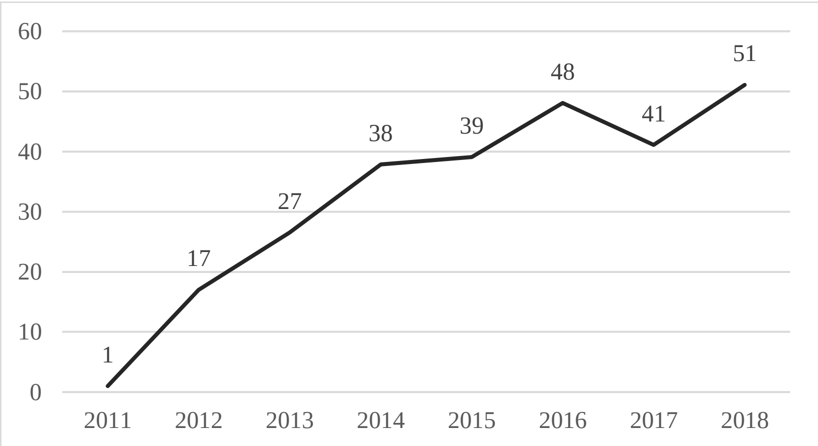

Figure 3. Trends of ‘Big data in Social Media’ searches from 2011 to 2018. (created by the authors, based on Google Trends data)

In the scientific literature, authors distinguish social media as one of the major data sources [11], [31], [41]. There are various definitions of social media provided by researchers, such as:

- Gruebner et al. [26] identify social media as online programmes that allow users to share and use content and communicate with each other in a variety of ways.
- Constantinides [18] writes that social media is defined as Web 2.0 applications that allow creating, editing and distributing user-generated content.
- Henderson and Bowley [29] defines social media as a set of programmes focused on user participation, information sharing and collaboration.
- According to Kaplan and Haenlein [32], social media is a group of web-based programmes that rely on the Web 2.0 ideological and technological bases to create and exchange user-generated content.
- Wildman and Obar [60] synthesised the definitions of various authors and identified four main features attributable to social media that are important for understanding the principles of social media performance: 1) these are Web 2.0 programmes; 2) these programmes are driven by user-generated content; 3) the principle of operation is based on individual user or group profiles; 4) by connecting individuals or groups, social media services facilitate the development of social networks. It can be noted that the definitions of the above-mentioned authors are also based on these features.

Digging deeper in the concept of social media, it is expedient to distinguish the types of social media. The main types of social media that are highlighted in literature are social networks, web blogs, microblogs, content communities etc. [11], [19], [33]. To gain a deeper understanding of social media, it is necessary to conduct an analysis of social media classification paradigms. One of the most popular classifications of social media can be Constantinides and Fountain's [19], a classification based on the use of social media. The five main categories of social media are distinguished into blogs, social networks, content communities, forums and content aggregators. Meanwhile, Kaplan and Haenlein [32] classify social media based on information that can be transmitted over a certain amount of time and self-disclosure – their classification of the above-mentioned social media forms is complemented by virtual social worlds and virtual gaming worlds. However, the purpose of this work is to focus on the use of social media data for marketing and communication purposes; the main types of social media remain social networks, news sites, content-sharing programmes, blogs and microblogs, forums and content aggregators.

Various types of social media differ in their purpose, functionality and generated data:

- Web blogs are a form of social media that is similar to a diary [54]. Blogs are defined as simple-to-use personal websites where users freely share their experiences, stories, content (texts and photos) or comment on other users' posts. With the Twitter application, the blog media has expanded the microblogging format.
- Microblogging allows you to publish small-scale text messages that are visible to all readers or a user-selected group of users [32].
- Social networks can be defined as a set of services that enable individual users to create personal profiles and share various information about themselves, publish pictures, write public announcements, search for other users, make new contacts and exchange contacts [17]. The popularity of social networks among consumers is growing rapidly. According to Statista [50], [51] data, in 2012 the average time spent on social networks per day was 90 minutes, but in 2017 this number increased to 135 minutes (Figure 4).

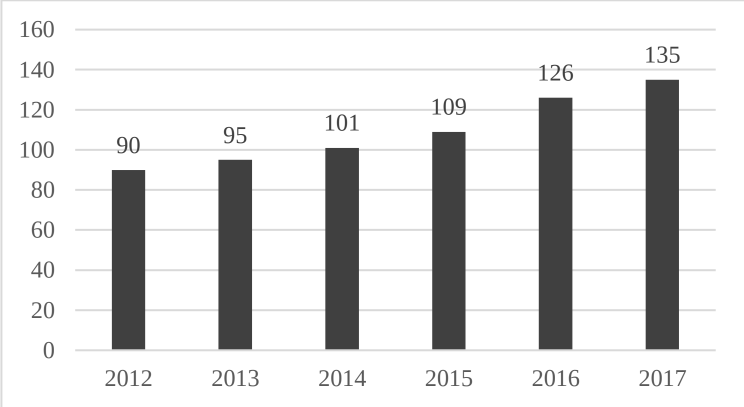

Figure 4. Daily time spent on social networks per day worldwide from 2012 to 2017. (created by authors, based on Statista [50], [51])

– Content communities are social media spaces where users share a particular type of information and content. The most popular content communities are pictures communities (e.g., Flickr or Picasa) and videos communities (e.g., YouTube, Vimeo).
– Forums are virtual discussion spaces, often related to specific themes or hobbies [19]. The content created by users in forums is easily found by search engines (e.g., Google) and accessible and visible for all internet users.
– Content aggregators – the first category of content aggregation systems includes programmes that allow users to easily access fully customised, syndicated web content using social bookmarking or RSS feeds; in the second category, content aggregators include programmes based on content gathered from multiple sources to create new, customisable content products or services (e.g., Google Maps) [18].

For more details on the types of social media, it can be noted that these types of social media are divided into smaller units – channels or platforms. For example, based on the data from the SimilarWeb platform, some of the most popular channels can be categorised as blogging platforms (Tumblr and Blogger), microblogging (Twitter and Tumblr; the platform has options to post both short messages and long texts), social networks (Facebook, LinkedIn and Google+), content communities (YouTube, Flickr and Instagram), forums (Quora and Stack Overflow), content aggregation applications (Flipboard and Alltop) and others. It is important to note that this is only a small part of the social media channels.

Zhu and Chen [62] classify social media based on the nature of the connections (profile-based or content-based) and the level of customisation of messages (broadcast or customised) – level of customisation describes the degree to which a service is customised to satisfy an individual's particular preferences. Social media (such as Facebook, Twitter and Whatsapp), and content-based social media can include content communities (such as Instagram, Flickr, Pinterest and YouTube). Based on the types of social media presented by the authors and the channels assigned to them, Figure 5 was created.

In summary, social media are Web 2.0-based programmes that enable users to create profiles, connect to communities and share information. It can be noted that social media have different forms of user involvement, channel specificity and type of data generated, covering both textual data (e.g., messages, comments and feedback), visual and complex data (e.g., videos and photos), interactions and operations. With the rapid growth of social media popularity and consumer engagement, data volumes are also growing creating favourable conditions for the application of social media data analysis in the business environment.

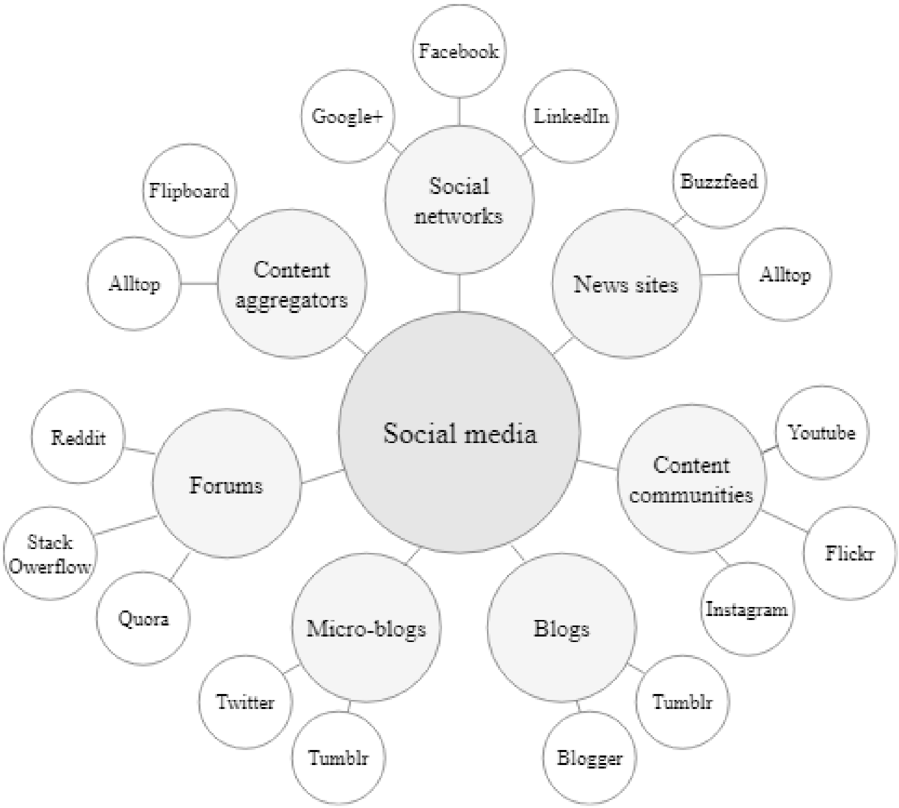

Figure 5. Social media classification (created by the authors, based on [19], [32]

5. Big data mining methods

It is important to emphasise that the value of data comes from collected data analytics. As for the analysis of social media data, Aggarwal [3] found two possible directions of social media data analysis: communication or structure analysis and content-based analysis. Relationship-based and structure-based analysis provides an opportunity to analyse network nodes, connections and communities and to identify developing regions. This analysis provides a good overview of core network behaviour, can be used to cognise the consumer or to predict a certain group's behaviour. Also, as social media contains different types of content, it is a favourable environment to apply textual and visual data analysis, which can provide valuable insights.

As mentioned earlier, big data has a complex structure and traditional analytical methods are not sufficient for them; that is why it is important to review possible techniques for machine learning analysis that can be applied for big data mining. Based on the analysis of scientific and methodological literature, it can be said that the general methods of machine learning data analysis include classification, clustering and regression [10]. These methods are divided into two types of machine learning methods: supervised and unsupervised methods. Supervised machine learning methods include classification and regression techniques, not supervised – the clustering method.

In most cases, the data retrieval method uses a classification method, whereby pre-categorised samples are used to prepare a classification model or algorithm [5]. The main task of this method is to create a data tagging algorithm by which the data will be divided into classes according to the submitted data. Data classification is a two-step approach [2]:

1. Learning – classification of available data by classification algorithm. At this stage, a classification model or algorithm is identified that analyses large amounts of data based on

the databases provided and their compatibility classes. This stage is considered as a learning phase because the model is based on the data from which the model or algorithm is learning.

2. Applying the method – this stage checks the accuracy and consistency of the model. Test data groups are analysed, and the accuracy of algorithm performance is checked – if it is accurate and the classification is done correctly, then the analysis of unknown and unmarked data is started.

The most popular and simple classification method is the decision tree method, but it is also possible to distinguish between support vector, neural network, Naive Bayes classification (based on Bayes theorem), association-based classification and methods of closest neighbours. In terms of social media data, clustering can be applied at the stage of data cleaning. Because the data are collected from many users and sources, a large number of data may be false or excessive (i.e., have extraneous data that are unnecessary or misleading), so it is necessary to process and prepare them before starting the data analysis. This can be carried out using a rule-based text classification or learning-based classification that works based on the analysis of tagged data [58].

Another method of machine learning applied to data mining is clustering. By applying this method, the algorithm looks for similarities between objects and when they are found, similar and different objects are grouped [38]. This is a self-contained, unmanaged data-purity technique that at first glance resembles a classification. Although these methods seem to be similar, they actually work based on different principles. By using the classification method, the data have assigned tags that assign them to classes. Using the clustering method, data are grouped according to similarities and differences within a data group, rather than based on common group characteristics (tags) [2].

For change and trend predictions, the machine learning regression method is used. Regression can also be used to model relationships between one or more variables and to detect change trends [44]. The following are the different types of regressions: linear regression, nonlinear regression, multidimensional linear regression and multidimensional nonlinear regression. The application of trend analysis includes forecasting the growth of customers or sales, predicting the effectiveness of advertising campaigns, remaining changes in consumer behaviour etc. [58].

To sum up, it can be noted that all the techniques of data retrieval are different – some of them require human intervention, others can act independently. Techniques also differ in the way in which they can return after the data analysis. Data tying techniques should be chosen based on the expected outcome and the desired outcome – whether the company seeks to identify common, recurring interests and differences between users or to identify changes. The company should choose a method that is suitable for analysing their data and helping to achieve their goals.

6. Possibilities of using big data generated by social media

Based on the analysed literature, the possibilities of using big data generated by social media are described briefly in the following sections.

6.1. Audience segmentation

For effective marketing, it is important to identify the audience we are planning to reach and influence. Social networking data allows us to identify relationships between users and detect different communities of customers who can share similar interests. Based on this data, companies can make decisions on how to adjust their strategy, adjust marketing strategies and predict how this might affect the future performance of a company. Also, while exploring relationships between users, it is possible to identify opinion leaders – influencers – who influence the community and use them in a communication strategy. The analysis of online consumer behaviour and online advertising audience segmentation – big data – provides the opportunity to monitor users' behaviour, where they are browsing, what are their interests and attribute to create a similar profile of consumer interest groups. This allowed advertisers to create and offer pre-designed segments of interest audience, and thus to advertise only to certain interested groups [20]. This provides an opportunity to reach consumers based not only on their demographic information (country, city, age and gender), but also on their behaviour or interests and attributes certain features to it, such as socio-economic class, family status and interests. For example, the online film company Netflix provides advertising messages that reflect its interests by analysing consumer behaviour and profile, which improves user's experience and engagement [1].

6.2. Micro-targeting

This is an audience-selection strategy used to create highly targeted ads. This strategy can be used for electoral or political campaigns [12]. A large amount of information about users is collected; by analysing the voters' habits – what parties they support, how often they vote etc. – and based on the collected data, it is possible to predict the election results, as well as to identify geographic areas where candidates should organise meetings with potential voters, activate promotional campaigns or personalise messages focusing on issues relevant to potential voter groups [14], [40]. Micro-targeting can also be applied in a business environment to reach a specific, narrow, niche audience. For example, when organising campaigns on Facebook platform, advertisers can divide the audience into very narrow niches equal to few people or one person [9].

6.3. User profiling

Social networks also allow setting up users' (customers') profiles. Based on the collected data, user profiles are created to identify user behavioural characteristics. User profiles are used in sales and marketing strategies [34]. The data generated by social networks can be used to create user profiles, as it provides an opportunity to track the user's public opinion and behaviour in the online environment [43]. Based on the user profile, the company can personalise promotional messages to potential customers [7]. User profiles can also be used for customisation. Clustering and classification methods are used for user segmentation or user profiling. Behavioural classification and decision tree provide an opportunity to determine what products can be presented to the consumer as a recommendation on what kind of reactions can be expected [34]. Based on consumer behaviour and interests, interest groups can be created to analyse consumer behaviour and make suggestions and messages to other users of similar behaviour. Also, based on the user profile, organisations can provide product or service recommendations, such as the e-commerce company ASOS or online television company Netflix.

6.4. Reputation management

When analysing text data, programmes can group data in real-time to allow the company to track real-time performance indicators and trends. One example of such software is Brand Fibers, which analyses real-time social media data. By discovering trademark-related messages, the programme is able to attribute them to the appropriate categories based on the data source (e.g., by referring to the social network where the comment was left and the link to the record) and visually displaying the brand's reputation changes and indicators. Methods of clustering and association analysis are the most commonly used methods of maintaining reputation management software [44].

Thus, big data provide the opportunity to get to know the client in real time, make better decisions, improve their performance and communicate more effectively. Choosing the right message for the right user and at the right time brings great value to the company and competitive advantage. Adapting messages is important because it establishes a closer relationship between the organisation and the consumer, which increases consumer loyalty, brand recognition and generates revenue and business opportunities [7]. Also, companies that know their user can use their advertising budget more efficiently, as the results of data analysis allow the company to focus on a segment of potential customers rather than random audiences, thereby saving advertising costs and ensuring more effective communication. In addition, it helps to improve the product or service according to the experience of the users, thus creating better relations with the user.

7. Challenges for big data projects

Big data is identified as a factor that can give the company a competitive advantage and added value. Despite the benefits of big data, in 2015, Gartner said that 60% of the big data projects would collapse. Before starting big data projects, it is important to understand what factors determine the failure of the project and what are the potential risks and challenges.

Big data has features such as volume, velocity and variety of data – it makes big data unique and valuable, but these features are also the cause of problems with data storage, analysis and visualisation ([31], [34]. Wamba et al. [55] named five potential difficulties for major data projects: legal aspects of data protection, technological and technical solutions, organisational culture (human resources), access to data and difficulties related to the industry's structure.

The rapidly growing data size requires the accumulation, efficient management and technological and technical resources. While current technologies are already able to store zetabytes of data – this amount is not enough, because data volumes are increasing so fast that it requires more and more storage space which can accommodate data. It is obvious that one of the resources that a company or organisation may lack are technical resources [48]. Before starting projects involving the collection and storage of big data, the company must foresee where the data will be collected – due to the high data growth rate, it is important to be able to increase the capacity of the database and have the financial resources to do so as the project launched can collapse due to excessive costs and lack of financial resources. Not only the data repositories, but also the software used to analyse the data, require significant financial resources.

Not only technical and financial resources, but human resources also have significant importance. It is important to ensure that employees who work with large amounts of data are sufficiently competent, as false insights can lead to the failure of the whole project. Thus, one more risk that can be distinguished from human activity is the competence of employees. In recent years, great attention has been paid to the protection of personal data and privacy. Companies that start with big data should provide where and for how long the data will be stored, how the data will be secured, ensure that the data will not leak, as well as the privacy and security of the individual will be ensured and personal data will not reach third parties [48].

Sivarajah et al. [49] distinguish three categories of data related to big data, based on the life cycle of data: data challenges, process challenges and management challenges. Based on the categories mentioned, the difficulties may be related to data and its characteristics (data size, variety, accuracy, detection, rendering, logical reasoning and relevance), processing and techniques (how data will be collected, stored, issues related to integration, analysis method selection and presentation of results). The difficulties encountered throughout the process are security, privacy, legal and ethnic aspects, management and resource allocation.

Gathering, processing and analysing big data are a complex process that requires a lot of resources (financial, technical and human resources). Before starting projects involving the collection and analysis of big data, the company should consider its capabilities and take a critical look at whether it really needs to initiate big data processes.

8. Conceptual framework of big data generated by social media usage

Based on the researches and analysis, a theoretical model for the use of big data generated by social media was created (Figure 6). The model is based on the principle of a cyclical model, as the completion of one project can be followed by repeated projects focusing on other possibilities of using social media data. The solid lines used in the figure indicate a direct link between the elements of the model and the sequence of actions. Dotted lines indicate a possible but optional sequence of actions.

Preparing for the development of a big data project requires complex planning that includes data collection, aggregation, analytical methods and further communication processes. Before starting the project, the organisation should define the purpose of the project and clear up what will be done during the data mining. This step is important because it narrows the data field by eliminating data that are not needed, thus reducing the possibility of false insights.

The second step is the analysis of internal resources. Following a literary analysis, it was noted that the lack of technological, financial and human resources can lead to a project collapse and potential risk factors; therefore, the organisation needs the necessary resources to ensure the smooth running of the project and prevent potential failures. Another important factor that should be considered at the beginning of the project is ensuring data security and the organisation's policy on the matter.

Social media data are made up of user-generated content and user interactions – these data can be obtained from social media platforms and software such as blogs, microblogging, social networks, forums, content aggregation systems and virtual social worlds. These sources may include unstructured data such as photos, videos, reviews, comments, various operations etc.

The data analysis process can be divided into three stages: data collection, aggregation and analysis. Because of the large amount of information and diversity of the large data, the analysis of such data requires advanced analytical techniques such as machine learning techniques, including classification, clustering and regression. When using the clustering and classification techniques, a user profile can be created. These methods can also be used for audience segmentation. With the exception of certain audience segments covering both broad and niche audiences, the organisation can conduct audience-oriented communication. By analysing collected data, valuable insights and trends can be identified, and this can lead to further communication strategy or require actions related to the reputation management of the organisation. It can also be a useful tool for evaluating the market position of an organisation.

The last step in this model is communication, which can be done through personalised messages or customised offers. In communication processes, consumer's responses to messages may return data, such as comments or feedback, in a form that determines it, and its process may become cyclical and repetitive from the beginning.

The model is theoretical and therefore requires empirical approval. The model illustrates the organisation's path from goal-setting to communication based on social media data analysis. As there is a lack of models about analytical methods and how companies should work with social media analytics, this model could be used to draw up an action plan before starting a major data-based communication project. With further research and literary analysis, the model can be complemented with technological analytics and data mining tools to create even greater value.

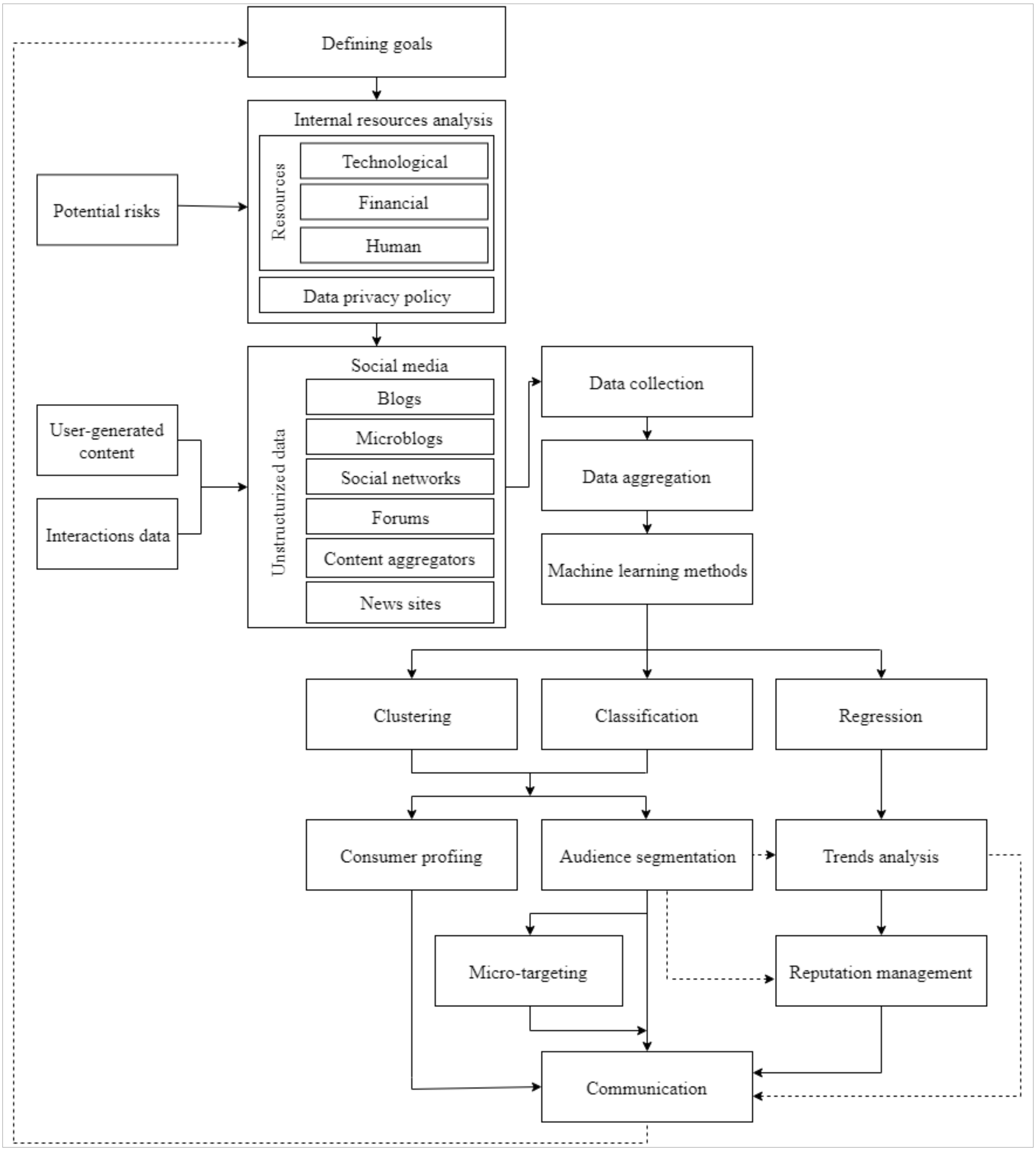

Figure 6. Conceptual framework of big data generated by social media usage.

9. Conclusion

An analysis of the scientific literature has shown that there is no single precise definition of big data. However, most authors define big data according to their attributes such as volume, velocity, variety and so on. Summing up the big data can be described as the volume of high-volume and high-speed information generated by people or devices requiring innovative and evolving technologies capable of collecting, storing and processing them. Big data is diverse and can be categorised based on different features, such as source, content, collection, storage and processing. The most important categories of data in the literature are structured, unstructured and semi-structured data, which are based on data content and format.

As the number of users of social media and social networks grows, the amount of data generated by social media is also increasing. This is useful for businesses as they can better understand their customers by collecting and analysing such data and discovering useful insights that can be used to adjust strategies, user profiling, audience segmentation or reputation management processes, or niche orientation to focus on narrower audiences.

Data from big data can be analysed in real-time using machine learning methods. In the analysis of social media data, classification and clustering methods are most commonly used. When planning big data projects, the company should define the goals of the project and choose an analytical method that is most effective for analysing collected data.

Many projects focused on big data fail due to insufficient technological, financial and human resources. Data analysis is a complex process – large amounts of data are difficult to manage, requiring high costs and expertise to get valuable insights from collected data. Before starting big data projects, companies should assess their position, capabilities and project goals, thus reducing project risk and eliminating unnecessary and unused data. Failure to assess the potential risks may result in inaccurate results, which can lead to project failure.

The model shows the types of social media and incoming data. The process of collecting, aggregating, analysing and utilising data from a big data analysis project is also illustrated. The model should help in planning the use of big data in marketing and communication strategies and in identifying potential risks and opportunities.

There is a lack of models in the literature on how companies should choose and apply different types of data mining. In this area, a deeper analysis of the possibilities of using social media data is needed, as well as research to identify the use of such data analysis methods in the market.

References

[1] Adweek. (2018). *How Marketing Teams Can Unleash Personalized Creative at Scale*. Accessed: Jan. 15, 2019. [Online]. Available: https://www.adweek.com/digital/how-marketing-teams-can-unleash-personalized-creative-at-scale/

[2] S. Agarwal, "Data mining: Data mining concepts and techniques," in *Proc. Int. Conf. Mach. Intell. Res. Advancement (ICMIRA)*, 2013, pp. 203–207, doi: 10.1109/ICMIRA.2013.45.

[3] C. C. Aggarwal, "An introduction to social network data analytics," in *Social Network Data Analytics*. Boston, MA, USA: Springer, 2011. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.3407&rep=rep1&type=pdf

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. Int. Conf. Web Search Data Mining*, 2008, pp. 183–193.

[5] A. B. E. D. Ahmed and I. S. Elaraby, "Data mining: A prediction for student's performance using classification method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014, doi: 10.13189/wjcat.2014.020203.

[6] A. Alharthi, V. Krotov, and M. Bowman, "Addressing barriers to big data," *Bus. Horizons*, vol. 60, no. 3, pp. 285–292, 2017, doi: 10.1016/j.bushor.2017.01.002.

[7] M. Anshari, M. N. Almunawar, S. A. Lim, and A. Al-Mudimigh, "Customer relationship management and big data enabled: Personalization & customization of services," *Appl. Comput. Inform.*, vol. 15, no. 2, pp. 94–101, 2019, doi: 10.1016/j.aci.2018.05.004.

[8] S. Banerjee and N. Agarwal, "Analyzing collective behavior from blogs using swarm intelligence," *Knowl. Inf. Syst.*, vol. 33, pp. 523–547, Jun. 2012, doi: 10.1007/s10115-012-0512-y.

[9] O. Barbu, "Advertising, microtargeting and social media," *Procedia-Social Behav. Sci.*, vol. 163, pp. 44–49, Dec. 2014, doi: 10.1016/j.sbspro.2014.12.284.

[10] B. Batrinca and P. C. Treleaven, “Social media analytics: A survey of techniques, tools and platforms,” *AI Soc.*, vol. 30, pp. 89–116, Jul. 2015, doi: 10.1007/s00146-014-0549-4.

[11] G. Bello-Orgaz, J. J. Jung, and D. Camacho, “Social big data: Recent achievements and new challenges,” *Inf. Fusion*, vol. 28, pp. 45–59, Mar. 2016, doi: 10.1016/j.inffus.2015.08.005.

[12] S. C. Boerman, “Political microtargeting: Relationship between personalized advertising on Facebook and voters' responses,” *Cyberpsychol., Behav., Social Netw.*, vol. 19, no. 6, pp. 367–372, 2016, doi: 10.1089/cyber.2015.0652.

[13] D. Boyd and K. Crawford, “Critical question for big data,” *Inf., Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012, doi: 10.1080/1369118X.2012.678878.

[14] H. Canick. (2016). *How Social Media, Microtargeting and Big Data Revolutionized Political Marketing*. Accessed: Jan. 21, 2019. [Online]. Available: https://www.ama.org/publications/MarketingNews/Pages/social-media-big-data-microtargeting-revolutionized-political-marketing.aspx?utm_content=buffer519bf&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

[15] J. M. Cavanillas, E. Curry, and W. Wahlster, Eds., *New Horizons for a Data-Driven Economy–-A Roadmap for Big Data in Europe*. 2015, doi: 10.1007/978-3-319-21569-3.

[16] Y. C. Chen, W. C. Peng, and S. Y. Lee, “Efficient algorithms for influence maximization in social networks,” *Knowl. Inf. Syst.*, vol. 33, no. 3, pp. 577–601, Dec. 2012, doi: 10.1007/s10115-012-0540-7.

[17] S. Čičević, A. Samčović, and M. Nešić, “Exploring college students' generational differences in Facebook usage,” *Comput. Hum. Behav.*, vol. 56, pp. 83–92, Mar. 2016, doi: 10.1016/j.chb.2015.11.034.

[18] E. Constantinides, “Foundations of social media marketing,” *Procedia-Social Behav. Sci.*, vol. 148, pp. 40–57, Aug. 2014, doi: 10.1016/j.sbspro.2014.07.016.

[19] E. Constantinides and S. J. Fountain, “Web 2.0: Conceptual foundations and marketing issues,” *J. Direct, Data Digit. Marketing Practice*, vol. 9, no. 3, pp. 231–244, 2008, doi: 10.1057/palgrave.dddmp.4350098.

[20] N. Couldry and J. Turow, “Advertising, big data and the clearance of the public realm,” *Int. J. Commun.*, vol. 8, no. 1, pp. 1710–1726, 2014, doi: 1932-8036/20140005.

[21] A. Cuzzocrea, I.-Y. Song, and K. C. Davis, “Analytics over large-scale multidimensional data: The big data revolution!” in *Proc. ACM 14th Int. Workshop Data Warehousing OLAP*, ACM, 2011, pp. 101–104.

[22] X. L. Dong and D. Srivastava, “Big data integration,” in *Proc. IEEE 29th Int. Conf. Data Engineering (ICDE)*, Apr. 2013, pp. 1245–1248, doi: 10.1109/ICDE.2013.6544914.

[23] 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, doi: 10.1016/j.ijinfomgt.2014.10.007.

[24] Gartner. (2015). *Gartner Says Business Intelligence and Analytics Leaders Must Focus on Mindsets and Culture to Kick Start Advanced Analytics*. Accessed: Sep. 23, 2018. [Online]. Available: https://www.gartner.com/newsroom/id/3130017

[25] P. Groves, B. Kayyali, D. Knott, and S. Kuiken Van, “The `big data' revolution in healthcare,” *McKinsey Quart.*, vol. 2, no. 3, 2013.

[26] O. Gruebner *et al.*, “Big data opportunities for social behavioral and mental health research,” *Social Sci. Med.*, vol. 189, no. 1, pp. 167–169, 2017. [Online]. Available: https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/26030/3/gruebner_ big_data_v-def.pdf

[27] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and U. Khan, “The rise of `big data' on cloud computing: Review and open research issues,” *Inf. Syst.*, vol. 47, pp. 98–115, Jan. 2015, doi: 10.1016/j.is.2014.07.006.

[28] W. He, S. Zha, and L. Li, “Social media competitive analysis and text mining?: A case study in the pizza industry,” *Int. J. Inf. Manage.*, vol. 33, no. 3, pp. 464–472, 2013, doi: 10.1016/j.ijinfomgt.2013.01.001.

[29] A. Henderson and R. Bowley, “Authentic dialogue? The role of `friendship' in a social media recruitment campaign,” *J. Commun. Manage.*, 2010, doi: 10.1108/13632541011064517.

[30] Internet World Stats. (2018). \emph[World Internet Usage and Population Statistics}. Accessed: Jan. 20, 2019. [Online]. Available: https://www.internetworldstats.com/stats.htm

[31] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, “Big data: Issues and challenges moving forward,” in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2014, pp. 995–1004, doi: 10.1109/HICSS.2013.645.

[32] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horizons*, vol. 53, no. 1,\break pp. 59–68, 2010, doi: 10.1016/j.bushor.2009.09.003.

[33] A. M. Kaplan and M. Haenlein, "Two hearts in three-quarter time: How to waltz the social media/viral marketing dance," *Bus. Horizons*, vol. 54, no. 3, pp. 253–263, 2011, doi: 10.1016/j.bushor.2011.01.006.

[34] A. A. Khade, "Performing customer behavior analysis using big data analytics," *Procedia Comput. Sci.*, vol. 79, pp. 986–992, Jan. 2016, doi: 10.1016/j.procs.2016.03.125.

[35] N. Khan *et al.*, "Big data: survey, technologies, opportunities, and challenges," *Sci. World J.*, vol. 2014, Jul. 2014, Art. no. 712826, doi: 10.1155/2014/712826.

[36] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, and R. Buyya, "The anatomy of big data computing," *Softw., Pract. Exper.*, vol. 46, no. 1, pp. 79–105, 2016, doi: 10.1002/spe.2374.

[37] J. Liu, J. Li, W. Li, and J. Wu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 134–142, May 2016, doi: 10.1016/j.isprsjprs.2015.11.006.

[38] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE 10th Int. Conf. Data Mining (ICDM)*, 2010, pp. 911–916, doi: 10.1109/ICDM.2010.35.

[39] A. McAfee *et al.*, "Big data: The management revolution," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 60–68, 2012.

[40] MIT Technology Review. (2012). *How Obama's Team Used Big Data to Rally Voters*. Accessed: Jan. 15, 2019. [Online]. Available: https://www.technologyreview.com/s/509026/how-obamas-team-used-big-data-to-rally-voters/

[41] E. Olshannikova, T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Conceptualizing big social data," *J. Big Data*, vol. 4, no. 1, p. 19, 2017, doi: 10.1186/s40537-017-0063-x.

[42] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, Oct. 2018, doi: 10.1016/j.jksuci.2017.06.001.

[43] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to Twitter user classification," in *Proc. ICWSM*, 2011, vol. 11, no. 1,\break pp. 281–288.

[44] S. Politaitė and J. Sabaitytė, "Didžiųjų duomenų naudojimas klientui pažinti," *Mokslas–Lietuvos Ateitis/Science–Future of Lithuania*, vol. *10*, pp. 1–10, 2018.

[45] V. Rajaraman, "Big data analytics," *Resonance*, vol. 21, pp. 695–716, Sep. 2016, doi: 10.1007/s12045-016-0376-7.

[46] P. Russom and T. Org, "Big Data Analytics," Tech. Rep., 2011.

[47] A. A. Senthilkumar, B. K. Rai, A. A. Meshram, and A. Gunasekaran, "Big data in healthcare management: A review of literature," *Amer. J. Theor. Appl. Bus.*, vol. 4, no. 2, pp. 57–69, 2018, doi: 10.11648/j.ajtab.20180402.14.

[48] S. Sharma, "Rise of big data and related issues," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2015, pp. 1–6, doi: 10.1109/INDICON.2015.7443346.

[49] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.

[50] Statista. (2017a). *Daily Time Spent on Social Networking by Internet Users Worldwide From 2012 to 2017 (in Minutes)*. Accessed: Jan. 7, 2019. [Online]. Available: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

[51] Statista. (2017b). *Number of Social Media Users Worldwide 2010–2021*. Accessed: Jan. 20, 2019. [Online]. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[52] Statista. (2018). *User-Generated Internet Content Per Minute 2018*. Accessed: Jan. 20, 2019. [Online]. Available: https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

[53] M. Tanwar, R. Duggal, and S. K. Khatri, "Unravelling unstructured data: A wealth of information in big data," in *Proc. 4th Int. Conf. Rel., Infocom Technol. Optim. (ICRITO), Trends Future Directions*, 2015, pp. 1–6, doi: 10.1109/ICRITO.2015.7359270.

[54] G. Thevenot, "Blogging as a social media," *Tourism Hospitality Res.*, vol. 7, nos. 3–4, pp. 287–289, 2007, doi: 10.1057/palgrave.thr.6050062.

[55] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study,” *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, Jul. 2015, doi: 10.1016/j.ijpe.2014.12.031.
[56] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li, “Unsupervised sentiment analysis for social media images,” in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 2378–2379, doi: 10.1109/ICDMW.2015.142.
[57] We Are Social. (2018). *Global Digital Report 2018*. [Online]. Available: https://digitalreport.wearesocial.com/
[58] F. Weiguo and M. D. Gordon, “Unveiling the power of social media analytics,” *Commun. ACM*, vol. 57, no. 6, 2014, doi: 10.1145/2602574.
[59] S. M. Weiss and N. Indurkhya, *Predictive Data Mining: A Practical Guide*. San Mateo, CA, USA: Morgan Kaufmann, 1998.
[60] J. A. Obar and S. Wildman, “Social media definition and the governance challenge: An introduction to the special issue,” *Telecommun. Policy*, vol. 39, no. 9, pp. 745–750, 2015, doi: 10.2139/ssrn.2637879.
[61] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 2, pp. 353–367, Jan. 2003, doi: 10.1109/TKDE.2013.109.
[62]] S. Yin and O. Kaynak, “Big data for modern industry: Challenges and trends [point of view],” *Proc. IEEE*, vol. 103, no. 2, pp. 143–146, Feb. 2015, doi: 10.1109/JPROC.2015.2388958.
[63] Y.-Q. Zhu and H.-G. Chen, “Social media and human need satisfaction: Implications for social media marketing,” *Bus. Horizons*, vol. 58, no. 3, pp. 335–345, 2015, doi: 10.1016/j.bushor.2015.01.006.