Social media analytics - Challenges in topic discovery, data collection, and data preparation

Stieglitz, Stefan; Mirbabaie, Milad; Roß, Björn; Neuberger, Christoph

DOI: http://dx.doi.org/10.1016/j.ijinfomgt.2017.12.002

URN: urn:nbn:de:hbz:464-20180112-120129-3

Link: http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=45142

Source: International Journal of Information Management, Volume 39, April 2018, Pages 156-168; available online 22 December 2017

International Journal of Information Management 39 (2018) 156–168

Contents lists available at ScienceDirect

International Journal of Information Management

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

Social media analytics – Challenges in topic discovery, data collection, and data preparation

Stefan Stieglitz[a,*], Milad Mirbabaie[a], Björn Ross[a], Christoph Neuberger[b]

[a] University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany
[b] Ludwig-Maximilians-Universität München, Oettingenstraße 67, 80538 München, Germany

ARTICLE INFO

Keywords:
Social media analytics
Social media
Information systems
Big data

ABSTRACT

Since an ever-increasing part of the population makes use of social media in their day-to-day lives, social media data is being analysed in many different disciplines. The social media analytics process involves four distinct steps, data discovery, collection, preparation, and analysis. While there is a great deal of literature on the challenges and difficulties involving specific data analysis methods, there hardly exists research on the stages of data discovery, collection, and preparation. To address this gap, we conducted an extended and structured literature analysis through which we identified challenges addressed and solutions proposed. The literature search revealed that the volume of data was most often cited as a challenge by researchers. In contrast, other categories have received less attention. Based on the results of the literature search, we discuss the most important challenges for researchers and present potential solutions. The findings are used to extend an existing framework on social media analytics. The article provides benefits for researchers and practitioners who wish to collect and analyse social media data.

1. Introduction

Social media has evolved over the last decade to become an important driver for acquiring and spreading information in different domains, such as business (Beier & Wagner, 2016), entertainment (Shen, Hock Chuan, & Cheng, 2016), science (Chen & Zhang, 2016), crisis management (Hiltz, Diaz, & Mark, 2011; Stieglitz, Bunker, Mirbabaie, & Ehnis, 2017a) and politics (Stieglitz & Dang-Xuan, 2013). One reason for the popularity of social media is the opportunity to receive or create and share public messages at low costs and ubiquitously. The enormous growth of social media usage has led to an increasing accumulation of data, which has been termed Social Media Big Data. Social media platforms offer many possibilities of data formats, including textual data, pictures, videos, sounds, and geolocations. Generally, this data can be divided into unstructured data and structured data (Baars & Kemper, 2008). In social networks, the textual content is an example of unstructured data, while the friend/follower relationship is an example of structured data.

The growth of social media usage opens up new opportunities for analysing several aspects of, and patterns in communication. For example, social media data can be analysed to gain insights into issues, trends, influential actors and other kinds of information. Golder and Macy (2011) analysed Twitter data to study how people's mood changes with time of day, weekday and season. In the field of Information Systems (IS), social media data is used to study questions such as the influence of network position on information diffusion (Susarla, Oh, & Tan, 2012).

Many existing research papers are isolated case studies (Kim, Choi, & Natali, 2016; Li & Huang, 2014; Oh, Hu, & Yang, 2016) that collect a large data set during a specific time frame on a specific subject and analyse it quantitatively. Despite the variety of disciplines such projects can be found in, they have much in common. The steps necessary to gain useful information or even knowledge out of social media are often similar. Therefore, the field of "Social Media Analytics" aims to combine, extend, and adapt methods for the analysis of social media data (Stieglitz, Dang-Xuan, Bruns, & Neuberger, 2014). It has gained considerable attention and subsequently acceptance in academic research, but there is still a lack of comprehensive discussions of social media analytics, and of general models and approaches. Aral, Dellarocas, and Godes (2013) presented a framework to organise social media research, and van Osch and Coursaris (2013) proposed a framework and research agenda explicitly limited to organisational social media. Both frameworks are geared towards classifying areas of research and, by extension, research questions, not methods to address these questions. While such frameworks are useful to decide what to research, and to locate individual projects within a larger context, they do not offer guidance

* Corresponding author.
E-mail address: stefan.stieglitz@uni-due.de (S. Stieglitz).

https://doi.org/10.1016/j.ijinfomgt.2017.12.002
Received 23 October 2017; Accepted 1 December 2017

on how to carry out the research, and which challenges might arise. Of course, there is also research that discusses challenges researchers face when employing specific methods for analysing social media data, such as social network analysis (Kane, Alavi, Labianca, & Borgatti, 2014) or opinion mining (Maynard, Bontcheva, & Rout, 2012), and there are literature reviews focused on specific goals such as the identification of users who are influential offline (Cossu, Labatut, & Dugué, 2016) or on specific topics such as social bots (Stieglitz, Brachten, Ross, & Jung, 2017b). Yet social media analytics consists of several steps, of which data analysis is only one. Before the data can be analysed, they have to be discovered, collected, and prepared. An overview of the challenges of social media analytics is needed to be able to manage the complexity of conducting social media analytics.

We therefore carried out a systematic literature review, arguing that the complexity of these equally important steps has not yet been adequately covered in research, and there are no widely accepted standards on how to proceed within each of the steps. We explicitly focused on papers that deal with the challenges researchers face when discovering topics, and when collecting and preparing social media data for analysis, regardless of the method they later use during the analysis.

Our paper focuses on the following research question:

- RQ: What challenges do researchers face when discovering topics, collecting and preparing social media data for further analyses?

The answers to this question will help researchers who have little experience with the analysis of social media data, and still be useful for those who are experienced. Newcomers to the field will find the overview of common challenges and proposed solutions useful, so that difficulties can be considered before they arise, when setting up the research design, instead of encountering problems in an advanced phase of the research. Experienced researchers will get a bird's eye view of the existing research, which helps identify areas that may need further investigation and challenges that have not been addressed adequately yet.

The remainder of our paper proceeds as follows: first we provide a status quo of the literature on social media analytics and highlight the theoretical background for our article afterwards. Second, we describe our research design and highlight our findings afterwards. Third, we discuss our results, point out their impact, and discuss a model for social media analytics. Finally, we conclude our article and derive aspects for further research.

2. Theoretical background

The interdisciplinary research field of social media analytics (SMA) deals with methods of analysing social media data. Researchers have divided the analytics process into several steps. We use the steps of discovery, collection, preparation, and analysis, which we adapted from Stieglitz et al. (2014). The particular challenges of social media data, however, have not been addressed comprehensively in the SMA literature. To be able to classify these challenges, we draw on theory from the big data literature instead. In particular, we use the four V's: volume, velocity, variety, and veracity.

2.1. Social media analytics

Since the rise of social media usage in the last decade, people have been seeking to gain information from the crowd as an additional source to traditional media. We use the term social media to refer to "Internet-based applications that build on the ideological and technological foundations of Web 2.0", where Web 2.0 means that "content and applications are no longer created and published by individuals, but instead are continuously modified by all users in a participatory and collaborative fashion" (Kaplan & Haenlein, 2010). Because of the broad definition of social media, its application purposes are manifold.

Despite the large variety of platforms, some characteristics are common to many of them. Because of the amount of the content produced daily and the number of active users on the platforms, organisations are motivated to understand which issues and trends evolve to identify risks and chances in the communication and derive useful implications. Besides the amount of content, it is also relevant for organisations to understand who creates the content and which actors are the most influential drivers in the communication. Both businesses and non-profit organisations seek to collect the data produced by the crowd in order to gain insights into mass communication. The data is often collected with tools which communicate with the respective API of the social media platform, if one exists, and crawl the data.

The term "Social Media Analytics" has gained a great deal of attention. It is defined as "*an emerging interdisciplinary research field that aims on combining, extending, and adapting methods for analysis of social media data*" (Zeng, Chen, Lusch, & Li, 2010). Whilst the perspective on the system is one important aspect, another aspect is the perspective on the users who create the content. Research that adopts this perspective explores different roles in the communication and the effects a respective role can have on the communication and the diffusion of information (Stieglitz et al., 2017c). Influencers or opinion leaders, for example, can be identified through a social network analysis, and by examining their follower network, one can reveal the reach of such an individual (Mirbabaie, Ehnis, Stieglitz, & Bunker, 2014; Mirbabaie & Zapatka, 2017). Furthermore, the behaviour of the roles is examined in order to understand the causes of a key role in the network and the effects it has on the overall network (Bhattacharya, Phan, & Airoldi, 2015; Kefi, Mlaiki, & Kalika, 2015; Mirbabaie et al., 2014; Zhang, Zhao, Lu, & Yang, 2016). Companies such as media agencies have recognised the importance of influencers and use them e.g. for product placement. Furthermore, the analysis of social media content evolved in the last few years to one of the main research purposes in Information Systems. One research goal might be to identify and analyse the information diffusion (Liu, 2015; Zhang & Zhang, 2016).

Among others, three domains in which social media is important and generates visible benefits are 1) in businesses, in 2) crisis communication, mainly in disaster management, and in 3) journalism and political communication.

In one of the main areas of social media analytics, businesses make use of social media data, for several purposes (Kleindienst, Pfleger, & Schoch, 2015). Social media data can be useful for detecting new trends in the communication or issues which could involve uncontrollable bad publicity (Bi, Zheng, & Liu, 2014). Social media is also used as a channel to communicate with customers (Griffiths & McLean, 2015; Pletikosa Cvijikj et al., 2013). For supporting decision-making processes, companies make use of social media reports, created ex post and based on predefined key performance indicators, or they make use of a dashboard for getting on-going analyses based on real-time social media data (Tsou et al., 2015). Social Media is also used for product placement (Liu, Chou, & Liao, 2015) in the social web.

Crisis communication research is an example of a field where social media data has had an impact. Social media is often used as a channel for emergency management agencies to inform people in an affected area on the current status of the respective crisis or how to behave (Liu, 2015). Social media data in the context of crisis communication can also be analysed to gain additional, previously unknown information, if volunteers e.g. take pictures or videos and spread the information into the crowd. Collected social media data can be also analysed for detecting a specific location or area where the crisis occurs. By analysing GPS data if it is included in the data or by applying the method of Named Entity Recognition the location could be also derived from the text (Alsudais & Corso, 2015; Bendler, Ratku, & Neumann, 2014; Mirbabaie, Tschampel, & Stieglitz, 2016). The spread of a disease can be monitored by mining emotional tweets (Ji, Chun, Wei, & Geller, 2015). Especially for Emergency Management Agencies, it is important to understand the communication behaviour and the current status

through social media, to be able to react faster and more efficiently. Furthermore, such agencies are also able to make use of the benefits of reaching a crowd through social media and diffuse relevant and life-saving information in their channels (Gill, Alam, & Eustace, 2014; van Gorp, Pogrebnyakov, & Maldonado, 2015).

Finally, social media platforms have been established in recent years as sources of data on political communication and for journalism. People debate on current issues and further actions of politicians and discuss the consequences. Social media analytics examines, for example, factors that influence political participation (Johannessen & Følstad, 2014; Meth, Lee, & Yang, 2015). Political parties and governments use social media as a channel to communicate with users, to reach a broader audience, in order to gain more followers on their political opinions (Blegind & Dyrby, 2013; Hofmann, 2014; Jungherr, Schoen, & Jürgens, 2016). People express their scepticism, fury, overall satisfaction or propose changes in social media. Through conducting social media analytics, governments and political parties are aiming to gain insights from the communication for deriving useful strategies for the next period of elections (Nulty, Theocharis, Popa, Parnet, & Benoit, 2016; Vaccari et al., 2013).

However, social media data can also have negative side effects (Wendling, Radisch, & Jacobzone, 2013). This has been recently labelled as "the dark side of social media" (Jalonen & Jussila, 2016; Kalhour & Ng, 2016; Payton & Conley, 2014). Rumours and false information could have a negative influence on the behaviour of other social media users. Therefore it becomes necessary to identify misinformation (Li, Sakamoto, & Chen, 2014; Wang, Ding, & Yang, 2014), rumours and fake news (Qin, Cai, & Wangchen, 2015), and the overall credibility of a user (Yu & Zou, 2015). Therefore, mechanisms are needed for detecting these categories of content. Another aspect is the usage of spam in social media data, which is not related to the topic and represents e.g. advertisement. Spam increases the amount of data and makes the analyses more difficult.

Overall, it can be stated that social media analytics is a highly complex process with different aspects regarding the respective application domain and the use of different methods. It is therefore useful and necessary to standardise this phenomenon to a process model, considering each step.

2.2. Steps of social media analytics

To explicate this process, researchers have developed frameworks that create a common basis for conducting social media analytics. Aral et al. (2013) describe research opportunities of social media analytics and propose a research framework for understanding the relationships among society, business, and social media. Their framework consists of four types of social media-related activities, and three levels of analysis that researchers may focus on when examining these activities. Similarly, in a review of the literature on organisational social media, van Osch and Coursaris (van Osch & Coursaris, 2013) classified relevant studies according the artefact, actor and activity they examined.

However, few research articles consider the steps of social media analytics. Such frameworks take the form of process models. Fan and Gordon (2014) propose a process for social media analytics consisting of three steps "capture", "understand", and "present". The authors state that the step of capture consists of gathering the data and preprocessing it, whereas pertinent information is extracted from the data in this step. Afterwards, noisy information, if existing in the data, should be removed. However, the core of this step consists of applying a key technique, such as a sentiment analysis or social network analysis, for understanding the data. In the last step the findings should be summarised and presented (Fan & Gordon, 2014). Stieglitz et al. (2014) also propose a framework for social media analytics (SMA), which is the most accepted one in information systems, based on the citations of the paper in IS literature. The authors describe the SMA process as consisting of three steps (see Fig. 1).

We adapt their framework, adding a discovery phase that comes before the tracking phase, for the following reasons. The framework was originally developed in the context of political communication. In principle, it can easily be adapted for other research domains. The goals and analysis methods might be different, but the process is essentially the same. The researchers still need to take the same decisions regarding data sources, approaches, software architecture and data storage. In politics, it is often known beforehand which topics should be tracked, e.g. the prevailing sentiment surrounding a political party. In a more general context the topics might not be known a priori, and have to be discovered first. Even when the topic on which data will be collected, such as a crisis situation, is already known, these methods can help identify the keywords and hashtags frequently used to talk about this topic. When employed as a preliminary step, this can help researchers achieve better coverage of a topic than would have been possible with terms defined a priori. Additionally, recent research has identified challenges commonly encountered in topic discovery (Chinnov, Kerschke, Meske, Stieglitz, & Trautmann, 2015). This suggests that the addition of this step and its explicit inclusion in a literature review results in a more comprehensive coverage of challenges.

This results in the following four-step framework:

- *Discovery*: The "uncovering of latent structures and patterns" (Chinnov et al., 2015)
- *Tracking*: This step involves decisions on the data source (e.g. Twitter, Facebook), approach, method and output. A detailed subdivision of this step can be found in Stieglitz et al. (2014). In several studies the completeness of different Twitter sources was compared (Driscoll & Walker, 2014; Morstatter, Pfeffer, & Liu, 2014; Morstatter, Pfeffer, Liu, & Carley, 2013).
- *Preparation*: Beyond this, the original framework does not elaborate on the preparation steps necessary.
- *Analysis*: Depending on the purpose there are several methods available, including social network analysis and opinion mining.

2.3. Types of challenges in big data analytics

As shown above, the existing SMA literature elaborates on the steps involved to some extent. However, to our knowledge, there is no comprehensive discussion of the challenges involved in these steps. To fill this gap, we draw on the literature on "big data". It can be argued that social media data shares many characteristics of "big" data, a term that encompasses data obtained from vastly different sources and in very different disciplines. It also includes nucleotide and protein sequences stored in massive bioinformatics databases (Howe et al., 2008) and weather and radar data used to predict flight arrival times (McAfee & Brynjolfsson, 2012). The two streams of research have much in common. Discussions of social media data are commonly found in publications on big data (Cao, Basoglu et al., 2015; McAfee & Brynjolfsson, 2012), and social media researchers frequently refer to the big data literature. This has been called "social big data" (Guellil & Boukhalfa, 2015) or "social media big data" (Lynn et al., 2015).

The notion that today's "big" data poses new challenges is widely acknowledged in various fields. The key factors by which this new phenomenon differs from traditional analytics can be summarised as follows:

- *volume*, the storage space required
- *velocity*, the speed of data creation coupled with the advantage gained from analysing the data in real time
- *variety*, the fact that data takes many different forms. It is often unstructured or its structure is specific to the data source, and
- *veracity*, uncertainty especially with regard to data quality.

The first three of these "four V's" were proposed by McAfee and Brynjolfsson (2012). Several other V's have been proposed in addition.

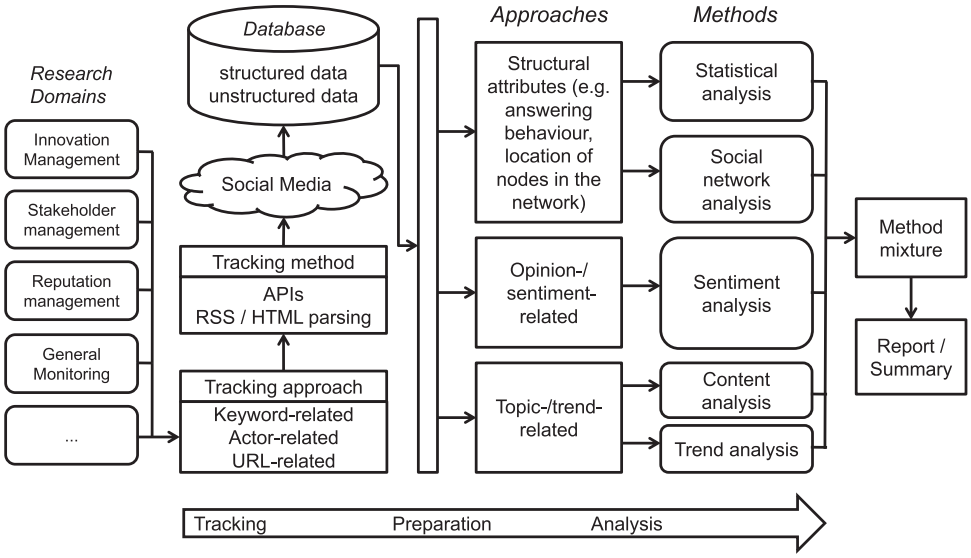

Fig. 1. The Social Media Analytics Framework (Stieglitz et al., 2014; Stieglitz & Dang-Xuan, 2013).

Veracity is frequently used. Some researchers use it only to refer to information security issues such as data integrity and authenticity (Demchenko, Grosso, Laat, & Membrey, 2013; Kepner et al., 2014). Others use a broader definition similar to the one given above (Artikis, Etzion, Feldman, & Fournier, 2012; Saha & Srivastava, 2014). Lukoianova and Rubin (2014) define the three dimensions of veracity as objectivity, truthfulness, and credibility. Another "V" sometimes proposed in the context of business analytics is *value* (Yin & Kaynak, 2015), which refers to the financial benefits generated by big data for an organisation. In the context of academic research, it is of course crucial that the research promises to be of value, but this is not a technical or methodological challenge.

Clearly the first four V's correspond to immediate technical challenges. For example, when the data takes up so much physical space that it does not fit into memory, many algorithms run considerably slower. The real-time nature and variety of the data directly influence architectural choices. boyd and Crawford (2012) argue that the use of big data in science raises methodological questions in addition to the technical ones. For example, data errors abound and must be dealt with, social media users are not representative of the general population, and publishing Facebook data is morally questionable when the data can easily be linked to individuals. Their concerns about ethics and access barriers are related to steps of the research process that are outside the scope of this article. Yet the data's lack of accuracy, representativeness and context is affected by the chosen data source and method of extraction. These issues fall under the broader definition of veracity.

In social sciences veracity is the main criterion for the assessment of big data (Bruns, 2013; King, 2011; Lin, 2015; Mahrt & Scharkow, 2013; Shah, Cappella, & Neuman, 2015). Social media promise a complete and real-time record of "natural" user activities. Issues relating to validity and representativeness have often been discussed and explored (Diaz, Gamon, Hofman, Kiciman, & Rothschild, 2016; Jungherr et al., 2016; Ruths & Pfeffer, 2014; Tufekci, 2014). It has even been debated and explored if SMA can replace traditional and more expensive ways of data collection such as population surveys (Diaz et al., 2016; Hargittai, 2015; Japec et al., 2015; Jungherr et al., 2016; Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016). But it was also criticised that there is a lack of tested standard procedures for data collection (Jungherr, 2016) and a danger of data-driven, non-theoretical approaches (Kitchin, 2014). We therefore use these four V's as categories for the purpose of classifying the individual difficulties faced by researchers. For example, spam and missing data both compromise the veracity of the data, and they are not likely to benefit from a technique that is designed to cope with its velocity. This classification allows us to determine quantitatively which types of problems are the most frequent, and which types of problems the proposed solutions address.

3. Research design

We chose to conduct a literature review to answer our research question. A review can "*tackle an emerging issue that would benefit from exposure to potential theoretical foundations*" (Webster & Watson, 2002). We argue that social media analytics is such an emerging research area that will benefit from a logical conceptualisation.

Our research design therefore consists of three principal steps. First, we use the theoretical foundations laid out above as a framework in classifying the existing research on the challenges of SMA. As Bem (1995) noted, "*a coherent review emerges only from a coherent conceptual structuring of the topic itself*". In our case, the steps of SMA and the challenges of big data serve as this conceptual a priori structure. This deductive step resulted in a rough categorisation of the articles found. In a second step, we examined the literature in more detail to identify similarities and differences between the individual articles. We thereby determined how the big data challenges become apparent in the SMA steps, and which solutions researchers have proposed. This step serves to inductively synthesise prior research and group related articles into logical concepts. Finally, in the third step, we considered the larger implications of our analysis for future research and derived an extension of the SMA framework.

Our literature review follows the systematic sequential process proposed by vom Brocke et al. (2009) and vom Brocke et al. (2015).

(1) First, we searched for predefined terms in the selected databases and read the title and abstract of each of the results to determine its relevance. The main problem we address in this article is which challenges researchers face when discovering, collecting and preparing social media data for further analysis. The search terms were chosen in order to identify papers from the area of social media analytics that explicitly mention challenges or difficulties. We expanded the search with other roughly synonymous search terms (see Table 1). We refined the search terms iteratively and formulated them so as to exclude many irrelevant publications but include many relevant ones. For example, we did not search for mentions of individual social media such as Twitter and Facebook because our aim was to uncover challenges that are common to many different platforms. Likewise, we limited the search to the title, abstract and keywords, which helped us only find articles that treat challenges as a crucial part of their content, and do not simply mention them as an afterthought, for example, when pointing out opportunities for future research. We considered

Table 1
Keywords and databases which were used for the Systematic Literature Review.

Search terms			Databases	Fields
("social media analytics"	AND	(challenges	ACM Digital Library	Title, Abstract
OR "social media analysis"		OR difficulties	AIS Electronic Library	
OR "social media data"		OR problems	IEEE Xplore	
OR "social media mining")		OR complexity)	ScienceDirect	

restricting our search further to include only articles that mentioned one of the SMA steps in the abstract. However, this would have greatly limited the number of articles considered because researchers may use other words for the steps, or not label them explicitly at all. Finally, we are aware that most of the search terms we used were coined fairly recently. Prior research into similar issues used other related terms such as Web 2.0 and User-Generated Content. Due to our choice of search terms, this research is not present in our review, and the oldest relevant paper is from 2011. We do not consider this restriction problematic because we aim to portray the state of the art, not the history of the field. Older problems that have not been solved yet are likely to be mentioned again in the current literature.

Our search was predominantly database-oriented and took into account all journal articles and conference publications from four bibliographic databases in order to represent the fields of computer science (ACM and IEEE), information systems (AIS) and the social sciences (ScienceDirect). The final search terms, databases and fields considered are listed in Table 1.

(2) To assess the potential relevance of a given hit, we carefully read the title, abstract and keywords. Relevant research publications are those that address challenges all researchers in SMA face during the discovery, collection and/or preparation phases, independent of the method they use later during data analysis. For example, publications were excluded if they only referred to challenges that are tied to specific methods, such as feature selection when using machine learning algorithms (Tang, Hu, Gao, & Liu, 2012), or difficulties associated with individual domains such as medical research (Wegrzyn-Wolska, Bougueroua, & Dziczkowski, 2011). Editorials and other non-research publications were also deemed irrelevant. We did not, however, exclude papers which demonstrate the feasibility of a solution in the context of a specific application if the underlying problem is likely to appear in other contexts as well. For example, Anderson et al. (2015) describe an architecture for the analysis of crisis-related social media data but their approach could equally be applied to any type of event.

(3) We then categorised papers that were relevant to our search according

- to the phase of the social media analytics process that the difficulties surfaced in (discovery, collection or preparation),
- and to the type of the problem (volume, velocity, variety or veracity).

Table 2 illustrates how we operationalised this categorisation, by providing example sentences from the classified articles that led to the corresponding categorisation.

(4) As the last step, we conducted a backward search, to find seminal highly cited papers which may also be relevant to the research question. To carry out the backward search we first collected all references from the relevant papers. Only references to other academic publications were counted. References to web pages, business reports and similar items were discarded. We then created a citation graph where each research article is a node and each citation an edge from the citing article to the cited article. The most frequently cited papers — the ones with the most incoming edges — can be assumed to be seminal publications that had a great deal of influence on the field. We determined their relevance according to the above criteria and read the relevant sections of the citing papers to determine the context they were frequently cited in.

4. Findings

4.1. Overview of the results

The execution of the systematic literature review, by searching for the search terms in all combinations and in all predefined databases and conducting a backward search, resulted in 49 relevant articles.

Table 3 shows the number of search results in each database. Of the articles returned by the search query, only about one in five were relevant to the research question. Most articles either dealt with the challenges of specific methods, such as feature extraction in machine learning, or domains, e.g. disaster response.

The classification enables us to take a closer look at the distribution of papers across categories (see Table 4). This makes it possible to examine which areas a large amount of research has been done in, and which ones have received less attention. This section is only intended as an overview of the current environment. We do not claim that areas in which less research has been done should receive more attention, because it may also simply mean that the problem is not as big as it may seem.

Challenges in the discovery step are most often due to the data volume. More precisely, the sheer volume of data is often cited as the primary motivation behind the development of topic discovery and event detection algorithms (Chang, Yamada, Ortega, & Liu, 2014; Chinnov et al., 2015; Hashimoto, Shepard, Kuboyama, & Shin, 2015). In contrast, there has been comparatively little research on discovery in a high-velocity, high-variety, or low-veracity environment. There are a few exceptions, however. Pletikosa Cvijikj and Michahelles (2011), who developed a trend detection system for Facebook, stress the importance of the real-time nature, or velocity, of social media, and Huang, Liu, and Nguyen (2015) mention the semi-structured nature of the data, or variety, as a challenge. Yang and Ng (2011) use an approach designed to cope specifically with noisy data, i.e. low veracity.

In the collection and preparation steps, data volume was also mentioned frequently. For example, Rehman, Weiler, and Scholl (2013) show how data warehousing can be extended to deal with social media data. However, variety was another challenge frequently mentioned, usually in relation to the processing of structured, semi-structured and unstructured data (Immonen et al., 2015).

Table 5 presents all the relevant papers found. Some of the categories are highly correlated. For example, all of the papers falling into the velocity category, i.e. streaming data, also mentioned the volume of data in one form or another.

Next, we present the results of the backward search. Recall that the purpose of the backward search was to find the most frequently cited publications that can be assumed to be seminal articles which greatly influenced the field. We therefore examined the most frequently cited publications more closely. The following Fig. 2 visualises the citation network.

Most of the highly cited articles deal with specific methods. For example, Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) is very widely used in the field of topic modelling. They all present event detection models and discuss challenges and difficulties which arise through data volume, velocity and variety. However, three of the publications would be considered relevant according to the criteria used earlier (Petrović et al., 2010; Ritter et al., 2012; Weng et al., 2011). All three of them motivate or discuss their models in the context of each of the challenges, from the ever increasing amount of real-time data to the variable and dynamic nature of the data and the noise in tweets in the form of "pointless 'babbles'" (Weng et al., 2011). However,

Table 2
Example mentions of steps and challenges in the articles.

	Example mention
Social media analytics step	
Discovery	"one of the challenges is to automatically tease apart the emerging topics of discussion from the constant background chatter" (Kasiviswanathan, Melville, Banerjee, & Sindhwani, 2011)
Collection	"It is challenging to collect social media data related to students' experiences because of the irregularity and diversity of the language used." (Chen, Vorvoreanu, & Madhavan, 2014)
Preparation	"performing a longitudinal analysis of these data becomes a Big-Data problem that cannot be tackled with traditional tools, storage or processing infrastructures" (Ruiz, Calleja, & Cazorla, 2015)
Challenge type	
Volume	"the sheer volume of the data produced on social media is overwhelming and acts as a major obstacle for manual inspection" (Khare, Torres, & Heravi, 2015)
Velocity	"we envision and develop a unified big data platform for social TV analytics, extracting valuable insights from TV social response in a real-time manner. Such a platform presents tremendous challenges …" (Hu, Wen, Gao, Chua, & Li, 2015)
Variety	"In the field of big data research, analytics on spatio-temporal data from social media is one of the fastest growing areas and poses a major challenge on research and application" (Zhang, Sun, Liu, Xu, & Wang, 2015)
Veracity	"the unstructured and uncertain nature of this kind of big data presents a new kind of challenge: how to evaluate the quality of data and manage the value of data within a big data architecture?" (Immonen, Paakkonen, & Ovaska, 2015)

Table 3
Number of search results per database.

Database	Number of results	Number of relevant results
ACM Digital Library	78	4
AIS Electronic Library	3	2
IEEE Xplore	122	34
ScienceDirect	46	6
Backward search	9	3
Total	260	49

Table 4
Number of search results by step and challenge.

Step	Volume	Velocity	Variety	Veracity
Discovery	15	5	4	3
Collection	19	11	17	7
Preparation	16	9	14	5

none of the publications found in the backward search deal specifically with data collection or preparation. It seems that researchers who face challenges in these areas have no widely accepted sources to turn to.

4.2. Identified major challenges and solutions

In addition to the classification of articles according to types of challenges, the following subsection provides an overview of how the challenges, such as data volume, manifest themselves in the individual steps. Articles are grouped under a common heading if they considered similar challenges at similar stages of the research process. This qualitative approach supplements the rough classification above, yet allows for a much more detailed analysis of the findings. In addition, we examined the articles to find possible solutions to the challenges, and summarise them. For example, data volume became apparent frequently in the discovery step, giving rise to the need for topic discovery and event detection algorithms, and the volume and velocity of data mean that in the collection phase, the choice of an appropriate software architecture becomes important.

4.2.1. Bridge the gulf between the social and the computational sciences

4.2.1.1. Challenge. Since social media analytics is an interdisciplinary field (Stieglitz et al., 2014), social media data is being analysed by researchers with very different backgrounds. Each discipline has its own tradition and merits, but also its own prejudices. In particular, Tinati et al. (2014), who consider social media analytics in a broader framework of Web Science, point out the gap between social and computer science and speak of an "unhelpful gulf". This gulf becomes evident throughout the entire research process, as social scientists do not have the methods at their disposal necessary to discover, collect and prepare relevant big social media data. On the other hand, many of the researchers who are currently applying computational approaches could benefit from a more solid grounding of their approaches in existing social theory.

4.2.1.2. Solutions. Tinati et al. (2014) argue that techniques from computer science and theories from social science should be combined to solve challenges in social media analytics. They propose a critical approach that makes use of social theory to ask critical questions early in the research process, before data collection, with practical implications on the decision which data should be collected. In the words of the authors, "Social action becomes an essential part of the data collection rather than only a product of analysis."

A possible solution that allows researchers to harness the power of social media big data while staying true to the established theories and methodologies from the social sciences is to integrate qualitative and quantitative research. For example, Chen, Vorvoreanu et al. (2014) examined issues and problems of engineering students based on their Twitter posts and saw the main challenge in the integration of qualitative analysis and large-scale data mining techniques. To solve this challenge, the authors first conducted a qualitative analysis of students' comments and then developed a multi-label classification algorithm on the basis of these results, in order to classify the tweets by the students and thus to gain insights into their needs and problems. Tinati et al. (2014) also present a system that helps researchers identify network roles in Twitter data by automatically extracting the data and calculating quantitative network metrics, but also enables an in-depth qualitative content analysis.

The information systems discipline can offer its own unique perspective in this area. IS researchers know that the call to approach social media data with a combination of wildly different methods from various disciplines, in particular, to combine qualitative and quantitative methods into a mixed-methods approach, is not new. This discipline has a long history of combining the two paradigms, and researchers struggling to accomplish the same in their own research may find the IS literature on the subject useful (Venkatesh, Brown, & Bala, 2013; Venkatesh, Brown, & Sullivan, 2016).

Finally, more broadly speaking, it seems from the analysed literature that the solution to this challenge is always to consider the big picture and think of new ways and different perspective to look at one's data. This requirement is of course not limited to research. As Carr et al. (2015) point out, in the business context, social media, which have traditionally mostly been used as a marketing research tool and to engage with the customers of an existing product, should also be considered in other areas such as product and category research.

Table 5
Relevant papers found in the Systematic Literature Review.

Database	Article	Phase			Challenge			
		Discovery	Collection	Preparation	Volume	Velocity	Variety	Veracity
IEEE	(Abbasi, Fu, Zeng, & Adjeroh, 2013)		X					X
IEEE	(Al-Qurishi et al., 2015)		X					
ACM	(Alsubaiee, Carey, & Li, 2015)			X	X	X		
IEEE	(Anderson et al., 2015)		X	X	X		X	
IEEE	(Bindra, Kandwal, Singh, & Khanna, 2012)		X					X
SD	(Cao, Wang et al., 2015)		X	X	X	X	X	
SD	(Carr et al., 2015)			X				X
SD	(Chae et al., 2014)		X	X	X		X	
IEEE	(Chang et al., 2014)	X			X			
IEEE	(Chen, Guo et al., 2014)			X			X	
IEEE	(Chen et al., 2016)			X			X	
IEEE	(Chen, Guo et al., 2014)		X	X	X			
AIS	(Chinnov et al., 2015)	X			X		X	X
IEEE	(Pletikosa Cvijikj and Michahelles, 2011)	X	X		X	X		
SD	(Garcia, 2013)			X	X			
IEEE	(Hashimoto et al., 2015)	X			X			
IEEE	(Hernandez et al., 2013)			X			X	
IEEE	(Hu et al., 2015)		X	X	X	X	X	
IEEE	(Huang, Dong, Yesha, & Zhou, 2014)	X	X	X		X	X	
IEEE	(Immonen et al., 2015)		X	X			X	X
ACM	(Kasiviswanathan et al., 2011)	X			X			
IEEE	(Kepner et al., 2014)		X	X				X
IEEE	(Khare et al., 2015)	X			X			
IEEE	(Kraft et al., 2013)		X	X	X	X		
IEEE	(Kumar & Rishi, 2015)		X		X	X	X	X
IEEE	(Lai, Rajashekar, & Rand, 2011)		X	X			X	
IEEE	(Li et al., 2014)	X			X			
IEEE	(Liu & Huet, 2012)	X			X			
IEEE	(Liu et al., 2012)		X	X			X	
IEEE	(Musaev, Wang, Shridhar, Lai, & Pu, 2015)			X	X	X		X
IEEE	(Patel, Gheewala, & Nagla, 2014)		X		X		X	
IEEE	(Qian, Zhang, Xu, & Shao, 2016)	X			X			
IEEE	(Rama Satish & Kavya, 2014)		X		X	X	X	X
IEEE	(Rehman, Weiler, & Scholl, 2013)			X	X	X	X	
ACM	(Reuter & Cimiano, 2012)	X			X			
IEEE	(Ruiz et al., 2015)		X		X		X	
IEEE	(Simmonds, Watson, Halliday, & Missier, 2014)		X	X	X	X		
IEEE	(Song & Kim, 2013)	X	X		X	X	X	
AIS	(Stieglitz et al., 2014)	X	X	X	X		X	X
ACM	(Tinati, Phillipe, Pope, Carr, & Halford, 2014)	X						
SD	(Vavliakis, Symeonidis, & Mitkas, 2013)	X			X	X		
IEEE	(Wang & Chen, 2015)		X		X	X	X	X
SD	(Weiler et al., 2016)	X	X	X	X	X		
IEEE	(Yang & Ng, 2011)	X			X			X
IEEE	(Zhang et al., 2015)		X	X	X		X	
IEEE	(Zhao et al., 2015)		X	X	X			
Backward search	(Petrović, Osborne, & Lavrenko, 2010)	X			X	X	X	X
	(Ritter, Etzioni, & Clark, 2012)	X			X	X	X	X
	(Weng, Yao, Leonardi, & Lee, 2011)	X			X	X	X	X

4.2.2. Discover relevant topics and events

4.2.2.1. Challenge. The volume of data makes it difficult to discover the relevant topics, trends and events in dynamic social media communication (Kasiviswanathan et al., 2011). However, the need for a structured and efficient approach is high across all of the application areas of social media analytics. Companies have realised the benefits of detecting the rise of new topics, such as discussions of their brands (Pletikosa Cvijikj & Michahelles, 2011). Likewise, for emergency response agencies, the quick detection of social media posts related to natural disasters is crucial (Kraft et al., 2013). Finally, social media have become a major news source for journalists (Khare et al., 2015).

4.2.2.2. Solutions. Most of the surveyed research focuses on developing topic modelling algorithms with the specific goal of event detection. While these algorithms are often tailored to the specific characteristics and unique challenges of social media data, they can usually be used with data from very different social media platforms. In this line of research, for example, Kasiviswanathan et al. (2011) state that the discovery of topics is a challenging task in the volume of social media data. They propose a solution for this challenge that makes use of dictionary learning, and evaluate it on Twitter data. Hashimoto et al. (2015) also evaluated their algorithm for detecting events accurately and in a short period of time on a Twitter data set. In their case the data set consisted of communication regarding an earthquake in Japan.

Event detection has also been attempted based on Flickr data (Liu & Huet, 2012; Reuter & Cimiano, 2012), making use of image tags and titles. In this research area, Liu and Huet (2012) propose an approach for detecting events in a given location that makes use of Latent Dirichlet Allocation (LDA). Reuter and Cimiano (2012) argue that the process of automating document management, i.e. "that incoming documents can be assigned to their corresponding event without any user intervention" is a challenge. They propose a system for classifying social media data into a set of events, which grows and evolves, and focus on Flickr as a data basis.

The argumentation of Qian et al. (2016) is that is has become "difficult to exactly find and organize the interesting events from

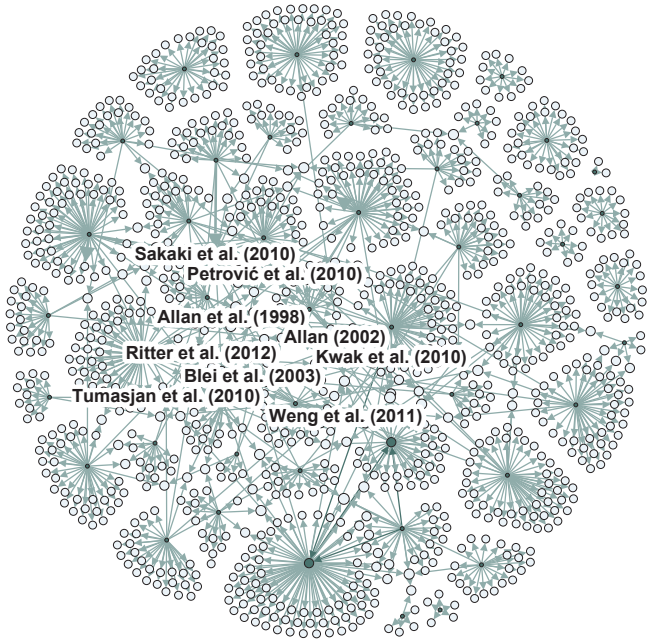

Fig. 2. Visualisation of the backward search results. Nodes represent the papers in the network and edges are citations. Green nodes are the papers which were found in the database search. The size of a node reflects the number of citations by other papers which were found in the database search. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

massive social media data." The authors go beyond the detection of topics and also consider the evolvement of trends in their research paper, by proposing a multi-modal social event tracking and evolution framework (mmETM), which is evaluated on Google News data. Finally, Vavliakis et al. (2013) also state that the identification of important events (online or real-life) from large textual documents is a problem. The authors propose an algorithm for detecting events semantically from document streams from social media, and evaluate in on blog posts.

In two cases, the algorithms were specifically tailored to the social media platform. Kraft et al. (2013) address the challenge of real-time detection of events on Twitter. They argue that the specific difficulty of Twitter data is the 140-character limit, because of which the tweet contents tend to contain very little of the information typically needed for reliable extraction. They propose methods to extract temporal and geographical indicators from tweets, and argue in favour of a visual approach for analysing events in real time. Pletikosa Cvijikj and Michahelles (2011) argue that the bulk of existing research addresses trend monitoring on Google and Twitter, whereas trend monitoring on Facebook is still challenging and under-researched. They suggest and evaluate a trend detection system based on the characteristics of content shared on Facebook. Thus, the authors develop 1) an algorithm for collecting data from Facebook and 2) an algorithm for trend detection over Facebook public posts. Building a real-time trend detection system for Facebook is not straightforward because Facebook's Graph API does not offer real-time streaming access in the way the Twitter Streaming API does. Therefore, their system relies on fetching the most recent posts repeatedly in short intervals to achieve near-real-time coverage.

Besides the research focused on developing new algorithms or modifying existing ones, there is also research on how to combine existing algorithms into frameworks for practical applications. Khare et al. (2015)'s framework for detecting events adopts an information retrieval perspective and is aimed at journalists.

In contrast to the above text-based methods for identifying individual topics and events, Chang et al. (2014) model the rise and fall of topics using a time series model, yielding insights into the way topics evolve over time. They argue that the reason buzz modelling is especially challenging is because their characteristic features, sudden spikes and heavy tails, are not captured by conventional time series models. They suggest a mixture of Product Life Cycle models as a solution, and they develop a probabilistic graphical model for discovering life cycle patterns in a collected dataset.

As it has become clear, there is already a large stream of research that deals with the detection of topics in social media data. However, as Chinnov et al. (2015) put it, "the need for automatic methods of topic discovery in the Internet grows exponentially with the amount of available textual information". We point the reader to their paper, which gives a more comprehensive overview of topic detection algorithms used in conjunction with social media data, and the challenges that arise there, and to the survey of event detection techniques presented by Goswami and Kumar (2016).

4.2.3. Choose an appropriate software architecture and storage technology

4.2.3.1. Challenge. The volume and velocity of data make it necessary to choose appropriate software architectures for the data collection stage. In conventional "small data" settings, a single machine, or a small group, runs a relational database management system (DBMS) that implements the Structured Query Language (SQL) standard, e.g. Microsoft SQL Server, PostgreSQL and MySQL. In the "big" data setting, these solutions are often no longer considered sufficient (Alsubaiee et al., 2015).

4.2.3.2. Solutions. Solutions specifically designed to handle social media "big" data (Kumar & Rishi, 2015; Patel et al., 2014) mostly focus on data storage technology and the algorithms used to process the data. In the "big" data setting, typical software architectures involve several layers of relational and non-relational database management systems, for permanent storage (such as Cassandra) and for caching (e.g. Redis), possibly additional technology for full-text search and real-time indexing such as Apache Solr (Anderson et al., 2015), and a web frontend to analyse and visualise the results.

Some of these architectures come from the tradition of data warehousing and online analytical processing (OLAP) technology, which is rooted in the field of business intelligence (Cao, Basoglu, Sheng, & Lowry, 2015; Liu et al., 2012). They involve a split between the transactional database, into which data is inserted at a high volume, and a separate database used only for analytical purposes. Implementations of data warehouses use a database schema, which formally describes the structure of the data, that is different from the ones used in conventional relational databases. Yet they build on many of the same technologies and rely on SQL. Moalla, Nabli, Bouzguenda, and Hammami (2017) provide a thorough review of data warehouse design for social media data.

Another frequently proposed solution for storing and processing large amounts of social media data is to make use of NoSQL. This umbrella term includes many different families of storage technologies that do not rely on relational schemas. Popular examples include Apache HBase (Huang et al., 2014), Cassandra (Anderson et al., 2015; Simmonds et al., 2014) and Redis (Song & Kim, 2013).

When the data is partitioned across several nodes in a computer cluster, new challenges arise with respect to how to process it. Efficient parallel implementations of algorithms are not always straightforward. The map-reduce paradigm is especially prominent among the solutions proposed for computational tasks such as the ones that arise in preprocessing. They often use the Apache Hadoop framework (Hu et al., 2015; Ruiz et al., 2015; Wang & Chen, 2015; Zhang et al., 2015; Zhao et al., 2015).

Writing a map-reduce job to analyse data can be significantly more difficult than writing a corresponding query for a single node, for which a single SQL statement is sometimes enough. Garcia (2013) describes how to implement a sequential algorithm in map-reduce, using an example from social media analytics. However, there are some attempts to reduce this burden, including Apache Pig (Anderson et al., 2015).

Programs are written in a conventional procedural style and then converted into a map-reduce job.

Yet, despite the proliferation of NoSQL-based solutions in current research, these technologies have drawbacks. Importantly they do not usually allow queries nearly as sophisticated as conventional SQL does. For example, Cassandra, another NoSQL database management system, does not efficiently provide a list of all row keys. After inserting Twitter data, if each tweet is stored as a row, simply listing all the tweets stored in the database is a time-consuming operation. Limiting this list to tweets published within a specific time frame which also contain one of a set of key words requires additional software development, as Cassandra does not natively provide this kind of functionality in an efficient way. Anderson et al. (2015) solved the problem by implementing an in-memory caching layer using Redis, which stores, for a list of tweets, the keys it can be found under in Cassandra. When Huang et al. (2014) designed a community discovery system using HBase, they equipped it with Apache Lucene to add full-text search capabilities, which HBase does not offer natively.

The previously mentioned articles address high-level architectural questions such as which database management system to use. Other articles relating to social media analytics found in the literature review were written by computer scientists who are developing new algorithms and data structures that power these database management systems under the hood, or who are rediscovering older technologies. Alsubaiee et al. (2015) describe the advantages of LSM-based indices in this context, and Kepner et al. (2014) address the question of what a database system should look like that preserves confidentiality.

4.2.4. Obtain high-quality data

4.2.4.1. Challenge. The veracity of data leads to issues in the data preparation step. The obtained social media data is often incomplete or noisy. Existing data may be of low quality. Apart from the problem of noisy and unreliable data, information may be missing altogether because the user did not choose to provide it, or because the financial or computational cost is too high to effectively collect it (Valkanas, Katakis, Gunopulos, & Stefanidis, 2014).

4.2.4.2. Solutions. To address the problem of low-quality data, one solution is to remove it by incorporating a filtering step in the preparation phase. To give an example from the analysed literature, Abbasi et al. (2013) faced medical sources of questionable credibility, and developed a crawler in response that filters out untrustworthy information.

To address the second problem of missing data, the naïve solution is to simply ignore incomplete observations. However, due to the amount of data missing, this may result in an undesirable reduction of the data set size. For example, researchers report consistently that only 1–2% of tweets contain global positioning system (GPS) coordinates (Hernandez et al., 2013; Valkanas et al., 2014). Ignoring all non-geotagged tweets may also lead to bias in the data, as only mobile devices typically add geolocations to tweets. Valkanas et al. (Valkanas et al., 2014) compared two differently sized samples of Twitter data, the smaller of which is the one available to the general public. They applied a number of popular analysis methods such as topic detection and sentiment analysis. For some research questions, they found the smaller sample adequate, but for others the information was stale and not representative.

The alternative solution for the problem of missing data is to infer it. Hernandez et al. (2013) use Twitter user descriptions to infer consumer profiles, predicting attributes such as parental status from the textual content. They also used textual clues such as “gotta love Florida football”, Foursquare check-ins, geo-tagged messages, time zone settings and mentions of regional events to infer user location with a precision of 94%. A few months’ worth of tweets was enough to infer these two attributes for more users than could be deduced from the profile description. Bindra et al. (2012) developed a method to correct for missing data in information cascade models. Musaev et al. (2015) present an event detection approach specifically geared towards dealing with noisy data and lack of geolocations.

4.2.5. Visualise the data meaningfully

4.2.5.1. Challenge. The volume and variety of data make it difficult to visualise the data in the preparation step. Visualisations can be crucial when decisions have to be taken quickly (Al-Qurishi et al., 2015; Liu et al., 2012). Decision makers, such as emergency management agencies, are forced to act quickly, and thus to save people’s lives. Social media data can support the decision-making process, as volunteers or bystanders share information about the crisis, but this is only possible when a clear and concise representation of the data can be found. This is especially difficult as the volume of data exceeds the capabilities of conventional tools, and the data to be visualised are available in different formats, e.g. textual content and geo-data.

4.2.5.2. Solutions. A considerable body of research is concerned with developing innovative solutions to the data visualisation problem. The focus of Yang and Ng (2011) lies on web opinion mining and its visualisation. They argue that the web opinions are short text passages and contain noisy content. Thus, classic document clustering techniques are inappropriate for clustering all documents. The authors suggest a density-based clustering algorithm and the scalable distance-based clustering technique for Web opinion clustering.

Other solutions address the visualisation of geographical information (Cao, Wang et al., 2015; Weiler et al., 2016). Cao, Wang et al. (2015) propose a general computational framework for dealing with geo-spatial social media data for scalable spatiotemporal analysis. The authors propose a data cube model for calculating the spatiotemporal distribution and dynamics and make use of the concept of space–time trajectories to visualise the activities of the users. The authors describe their implementation of the framework using Twitter as the main data source. Weiler et al. (2016) also address the visualisation of social media data in their work, with respect to the issues of detecting events and topics. However, the authors also consider the spatial and temporal dimension in the visualisation. The main data source for their research paper is Twitter. The authors address the issue by suggesting a clock-face metaphor solution and visualising, besides the spatial and temporal dimension, also the sentiments of the content.

Many of the articles found in our literature review make explicit reference to the research field of visual analytics (Chae et al., 2014; Chen, Guo et al., 2014; Chen et al., 2016). Visual analytics integrates interactive visualisations and the automated processing of data (Keim et al., 2008). Data visualisations are not only considered as an output of the research process, or as a way of communication results, but as an integral part of it. One of the stated goals of this discipline is to “derive insight from massive, dynamic, ambiguous and often conflicting data” (Keim et al., 2008). Chen, Vorvoreanu et al. (2014) propose a visual analytics system for analysing the public (social media) behaviour. Their system is built for disaster management and evacuation planning and supports decision makers to verify and examine certain aspects of the crisis situations, by considering spatial and temporal data. et al. (2014) also propose in their work a visual analytics tool for detecting patterns in people’s daily lives, i.e. the geolocations, by using an interactive multi-filter visualisation approach. Because of the sparseness and irregularity type of the data, the authors propose a self-developed system, track the movements of the users and analyse these in their system. As the data source, the authors used Weibo, a Chinese microblogging service.

5. Discussion

In this article, we set out to summarise the most frequently mentioned challenges researchers face even before they can begin to analyse social media data. As the results show, the discovery, collection and preparation of social media data is no easy task. There are plenty of

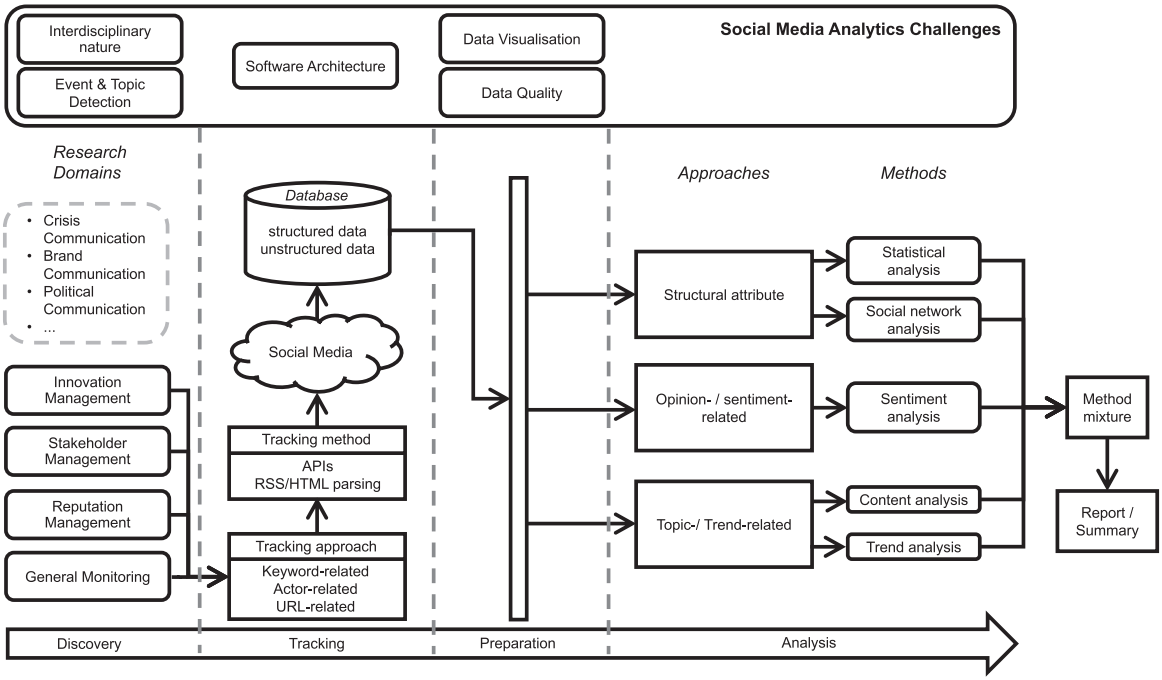

Fig. 3. The identified challenges in the context of the Social Media Analytics Framework (Stieglitz et al., 2014; Stieglitz & Dang-Xuan, 2013).

challenges to be encountered in each of the steps, and they need to be addressed adequately. Fig. 3 visualises our findings in the context of the original framework of Stieglitz et al. (2014). The challenges identified are placed above the steps they are most likely to arise in. To ensure the success of a new social media analytics project, researchers and practitioners should plan ahead and carefully consider how they will address each of these challenges well before they arise.

We drew on the four V's from the big data literature to categorise the challenges found in our analysis of the literature. This approach allowed us to show that the volume of data is the most frequently mentioned challenge overall. Researchers seem to feel inundated with the overabundance of social media data. This intimidating effect looks to be the strongest in the early stages of a research endeavour, since the challenge of volume was found to be especially predominant in the discovery phase. Not all of the data is relevant to the research topic, and thus irrelevant topics need to be filtered out. Advanced topic detection algorithms promise to solve this problem.

In later stages, the variety of data becomes another major challenge. The dynamic nature of social media data makes its collection and preparation for analysis especially complicated. Through literature search, we identified solutions from sophisticated software architectures to visual analytics.

Topic discovery and event detection are already well-established research fields. This is demonstrated by the concentration of citations observed during the backward search. Many of the articles on these topics cite a few high-profile publications around which research is centred. The same cannot be said for the other two stages, data collection and preparation. The backward search revealed no high-profile publications on these stages. Existing publications on these subjects do not seem to reach a wide enough audience. Yet, more individual papers mention challenges in the first two stages than in the discovery phase. This finding clearly emphasises the need for more articles addressing the early stages of the research process. In identifying the individual challenges researchers encounter in these stages, and pointing researchers to other relevant articles, we contribute to filling this gap.

In the papers that already document the tracking and preparation steps, usually in the research methodology section, these steps are often dealt with superficially, whereas a much longer portion of the section is devoted to data analysis. If the documentation is lacking and there are no standardised procedures and published best practices, research becomes more difficult to reproduce and access to the field is more difficult for researchers without a technical background. This problem mirrors the divide in the research community between social scientists and computer scientists, which was revealed in the literature review. boyd and Crawford (2012) highlighted that researchers are divided into the "data rich" and the "data poor". This concerns the financial means available to universities, since access to data may depend on funding. However, it also concerns skill sets: Researchers with a technical background are more likely to be able to collect and analyse "big data". The value of social science is not always recognised any more, and the resulting perceived hierarchy is problematic. The database search revealed similar results. Tinati et al. (2014) stress the "methodological impasse" that social science offers a rich theoretical understanding of human social interactions, but lacks the expertise to deal with the scale and dynamism of real social media data—in other words, its volume and variety. Meanwhile the sophisticated computational approaches which have been developed to deal with these challenges have a tendency to give little weight to the social nature of the data, limiting themselves to a technical perspective.

As conventional tools such as SPSS and Excel fail when used with datasets of several million rows, more and more researchers can benefit from at least a cursory understanding of programming, which enables them to quickly run analyses without the help of software developers. In many domains, researchers are increasingly turning towards languages such as R that solve precisely this problem. Therefore, in order to bridge the divide between those with technical skills and those without, we call upon researchers to document their data collection process more thoroughly so it can be replicated more easily by others.

Of course, any literature review comes with certain limitations. In this case, we relied on the papers specifically mentioning the challenges they solve. Some authors may have proposed solutions without giving examples of applications. In addition, the fact that many researchers propose solutions to a problem does not necessarily mean that many researchers face this problem in the first place.

6. Conclusion

Social media analytics is still a relatively new research area, but it is

of great interest to the Information Systems community and many researchers are embarking on SMA projects in our field. This article contributes to the Information Systems literature by presenting a summary of the main challenges and difficulties researchers face in the steps of the social media analytics research process that come before the data is analysed: discovery, collection and preparation. As a second contribution to the literature, we also point researchers to possible solutions for these challenges. These findings are equally relevant to practitioners, as businesses are increasingly looking to extract meaningful information from social media data, and are facing many of the same challenges researchers do.

Conceptualising the problem using the three-step social media analytics framework by Stieglitz et al. (2014) and the four "big data" V's provides a framework in which to think about possible difficulties before they arise. Which volume of data do we expect? How do we discover the parts which are relevant to our research? Do we have adequate infrastructure to cope with that volume when collecting and preparing the data? Which format will the data be in? If the data is unstructured, how can we extract the relevant structured information from it? This article is meant to help researchers ask and find answers to questions such as these. If the challenges highlighted above are addressed successfully, the social media analytics project will be much more likely to be a success.

References

Abbasi, A., Fu, T., Zeng, D., & Adjeroh, D. (2013). Crawling credible online medical sentiments for social intelligence. *Proceedings – SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013,* 254–263. http://dx.doi.org/10.1109/SocialCom.2013.43.

Al-Qurishi, M., Al-Rakhami, M., Alrubaian, M., Alarifi, A., Rahman, S. M. M., & Alamri, A. (2015). Selecting the best open source tools for collecting and visualizing social media content. *2015 2nd world symposium on web applications and networking (WSWAN),* 1–6. http://dx.doi.org/10.1109/WSWAN.2015.7210346.

Alsubaiee, S., Carey, M. J., & Li, C. (2015). *LSM-Based storage and indexing: an old idea with timely benefits. Second international ACM workshop on managing and mining enriched geo-spatial data*. New York, NY, USA: ACM1–6. http://dx.doi.org/10.1145/2786006.2786007.

Alsudais, K., & Corso, A. (2015). GIS, big data, and a tweet corpus operationalized via natural language processing. *AMCIS 2015 proceedings*.

Anderson, K. M., Aydin, A. A., Barrenechea, M., Cardenas, A., Hakeem, M., & Jambi, S. (2015). Design Challenges/Solutions for environments supporting the analysis of social media data in crisis informatics research. *Proceedings of the annual Hawaii international conference on system sciences,* 163–172. http://dx.doi.org/10.1109/HICSS.2015.29.

Aral, S., Dellarocas, C., & Godes, D. (2013). Introduction to the special Issue—Social media and business transformation: a framework for research. *Information Systems Research, 24*(1), 3–13. http://dx.doi.org/10.1287/isre.1120.0470.

Artikis, A., Etzion, O., Feldman, Z., & Fournier, F. (2012). Event processing under uncertainty. *Proceedings of the 6th ACM international conference on distributed event-based systems (DEBS '12),* 32–43. http://dx.doi.org/10.1145/2335484.2335488.

Baars, H., & Kemper, H.-G. (2008). Management support with structured and unstructured data – An integrated business intelligence framework. *Information Systems Management, 25*(2), 132–148. http://dx.doi.org/10.1080/10580530801941058.

Beier, M., & Wagner, K. (2016). Social media adoption: barriers to the strategic use of social media in SMEs. *Proceedings of the european conference on information systems*.

Bem, D. J. (1995). Writing a review article for Psychological Bulletin. *Psychological Bulletin, 118*(2), 172–177. http://dx.doi.org/10.1037/0033-2909.118.2.172.

Bendler, J., Ratku, A., & Neumann, D. (2014). Crime mapping through geo-spatial social media activity. *Proceedings of the international conference on information systems*.

Bhattacharya, P., Phan, T., & Airoldi, E. (2015). Investigating the impact of network effects on content generation: Evidence from a large online student network. *Proceedings of the international conference on information systems*.

Bi, G., Zheng, B., & Liu, H. (2014). Secondary crisis communication on social media: The role of corporate response and social influence in product-harm crisis. *PACIS 2014 proceedings*.

Bindra, G. S., Kandwal, K. K., Singh, P. K., & Khanna, S. (2012). Tracing information flow and analyzing the effects of incomplete data in social media. *2012 Fourth international conference on computational intelligence, communication systems and networks,* 235–240. http://dx.doi.org/10.1109/CICSyN.2012.51.

Blegind, T., & Dyrby, S. (2013). Exploring affordance of facebook as a social media platform in political campaigning. *Proceedings of the european conference on information systems*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society, 15*(5), 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878.

Bruns, A. (2013). Faster than the speed of print: Reconciling 'data' social media analysis and academic scholarship. *First Monday, 18*(10), s1. http://dx.doi.org/10.5210/fm.v18i10.4879.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems, 51*, 70–82. http://dx.doi.org/10.1016/j.compenvurbsys.2015.01.002.

Cao, J., Basoglu, K. A., Sheng, H., & Lowry, P. B. (2015). A systematic review of social networks research in information systems: Building a foundation for exciting future research. *Communications of the Association for Information Systems, 36*(January), 727–758.

Carr, J., Decreton, L., Qin, W., Rojas, B., Rossochacki, T., & Wen Yang, Y. (2015). Social media in product development. *Food Quality and Preference, 40*(Part B), 354–364. http://dx.doi.org/10.1016/j.foodqual.2014.04.001.

Chae, J., Thom, D., Jang, Y., Kim, S., Ertl, T., & Ebert, D. S. (2014). Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics, 38*, 51–60. http://dx.doi.org/10.1016/j.cag.2013.10.008.

Chang, Y., Yamada, M., Ortega, A., & Liu, Y. (2014). Ups and downs in buzzes: life cycle modeling for temporal pattern discovery. *2014 IEEE International Conference on Data Mining (ICDM),* 749–754. http://dx.doi.org/10.1109/ICDM.2014.28.

Chen, F., & Zhang, L. (2016). How to integrate social media in IS curriculum, especially for a small IS program? *Proceedings of the americas conference on information systems*.

Chen, S., Guo, C., Yuan, X., Zhang, J., & Zhang, X. L. (2014). MovementFinder: Visual analytics of origin-destination patterns from geo-tagged social media. *2014 IEEE conference on visual analytics science and technology, VAST 2014 – proceedings,* 239–240. http://dx.doi.org/10.1109/VAST.2014.7042509.

Chen, X., Vorvoreanu, M., & Madhavan, K. P. C. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies, 7*(3), 246–259. http://dx.doi.org/10.1109/TLT.2013.2296520.

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., ... Zhang, J. (2016). Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 270–279. http://dx.doi.org/10.1109/TVCG.2015.2467619.

Chinnov, A., Kerschke, P., Meske, C., Stieglitz, S., & Trautmann, H. (2015). An Overview of Topic Discovery in Twitter Communication through Social Media Analytics. *Proceedings of the americas conference on information systems*.

Cossu, J.-V., Labatut, V., & Dugué, N. (2016). A review of features for the discrimination of twitter users: Application to the prediction of offline influence. *Social Network Analysis and Mining, 6*(1), http://dx.doi.org/10.1007/s13278-016-0329-x.

Pletikosa Cvijikj, I., & Michahelles, F. (2011). Monitoring trends on Facebook. *Proceedings – IEEE 9th international conference on dependable, autonomic and secure computing, DASC 2011,* 895–902. http://dx.doi.org/10.1109/DASC.2011.150.

Demchenko, Y., Grosso, P., Laat, C., & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure. *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013,* 48–55. http://dx.doi.org/10.1109/CTS.2013.6567203.

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLoS One, 11*(1), http://dx.doi.org/10.1371/journal.pone.0145406.

Driscoll, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication, 8*(1), 1745–1764.

Fan, W., & Gordon, M. D. (2014). The Power of Social Media Analytics. *Commun. ACM, 57*(6), 74–81. http://dx.doi.org/10.1145/2602574.

Garcia, C. (2013). Demystifying mapreduce. *Procedia Computer Science, 20*, 484–489. http://dx.doi.org/10.1016/j.procs.2013.09.307.

Gill, A., Alam, S., & Eustace, J. (2014). Using Social Architecture to Analyzing Online Social Network Use in Emergency Management. In Proceedings of the Americas Conference on Information Systems.

Golder, S. A., & Macy, M. W. (2011). Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science, 333(6051), 1878—1881. https://doi.org/10.1126/science.1202775.

Goswami, A., & Kumar, A. (2016). A survey of event detection techniques in online social networks (No. 1). *(Keine Angabe)*, 6.

Griffiths, M., & McLean, R. (2015). Unleashing corporate communications via social media: A UK study of brand management and conversations with customers. *Journal of Customer Behaviour, 14*(2), 147–162. http://dx.doi.org/10.1362/147539215X14373846805789.

Guellil, I., & Boukhalfa, K. (2015). Social big data mining: A survey focused on opinion mining and sentiments analysis. *12th international symposium on programming and systems, ISPS 2015,* 132–141. http://dx.doi.org/10.1109/ISPS.2015.7244976.

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 63–76. http://dx.doi.org/10.1177/0002716215570866.

Hashimoto, T., Shepard, D., Kuboyama, T., Shin, K. (2015). Event Detection from Millions of Tweets related to the Great East Japan Earthquake using Feature Selection Technique. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (Vol. 1, pp. 7–12). doi: https://doi.org/10.1109/ICDMW.2015.248.

Hernandez, M., Hildrum, K., Jain, P., Wagle, R., Alexe, B., Krishnamurthy, R., ... Venkatramani, C. (2013). Constructing consumer profiles from social media data. *Proceedings – 2013 IEEE international conference on big data, big data 2013,* 710–716. http://dx.doi.org/10.1109/BigData.2013.6691641.

Hiltz, S.R., Diaz, P., & Mark, G. (2011). Social media and collaborative systems for crisis management, ACM Transactions on Computer-Human Interaction, Vol. 18 Issue 4, December 2011, ACM New York, NY, USA, DOI: https://doi.org/10.1145/2063231.

2063232.
Hofmann, S. (2014). Just Because We Can – Governments' Rationale for Using Social Media. *Proceedings of the European Conference on Information Systems*.
Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature, 455*(7209), 47–50. http://dx.doi.org/10.1038/455047a.
Hu, H., Wen, Y., Gao, Y., Chua, T. S., & Li, X. (2015). Toward an SDN-enabled big data platform for social TV analytics. *IEEE Network, 29*(5), 43–49. http://dx.doi.org/10.1109/MNET.2015.7293304.
Huang, Y., Liu, Z., & Nguyen, P. (2015). Location-based event search in social texts. *2015 International Conference on Computing, Networking and Communications, ICNC 2015,* 668–672. http://dx.doi.org/10.1109/ICCNC.2015.7069425.
Huang, Y., Dong, H., Yesha, Y., & Zhou, S. (2014). A scalable system for community discovery in twitter during hurricane sandy. *Proceedings – 14th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2014,* 893–899. http://dx.doi.org/10.1109/CCGrid.2014.122.
Immonen, A., Paakkonen, P., & Ovaska, E. (2015). Evaluating the Quality of Social Media Data in Big Data Architecture. *IEEE Access, 3*, 2028–2043. http://dx.doi.org/10.1109/ACCESS.2015.2490723.
Jalonen, H., & Jussila, J. (2016). Developing a conceptual model for the relationship between social media behavior, negative consumer emotions and brand disloyalty. *Lecture Notes in Computer Science, 9844*, 134–145. http://dx.doi.org/10.1007/978-3-319-45234-0_13.
Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., ... Usher, A. (2015). Big Data in Survey Research. *Public Opinion Quarterly, 79*(4), 839–880. http://dx.doi.org/10.1093/poq/nfv039.
Ji, X., Chun, S. A., Wei, Z., & Geller, J. (2015). Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining, 5*(1), 1–25. http://dx.doi.org/10.1007/s13278-015-0253-5.
Johannessen, M. R., & Følstad, A. (2014). Political social media sites as public sphere: A case study of the norwegian labour party. *Communications of the Association for Information Systems, 34*(1), 1067–1096.
Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. Journal of Information Technology & *Politics, 13*(1), 72–91. http://dx.doi.org/10.1080/19331681.2015.1132401.
Jungherr, A., Schoen, H., & Jürgens, P. (2016). The Mediation of Politics through Twitter: An Analysis of Messages posted during the Campaign for the German Federal Election 2013. *Journal of Computer-Mediated Communication, 21*(1), 50–68. http://dx.doi.org/10.1111/jcc4.12143.
Kalhour, M., & Ng, J. C. Y. (2016). The dark side of social media game: The addiction of social gamers. *Economia e Politica Industriale, 43*(2), 219–230. http://dx.doi.org/10.1007/s40812-016-0025-x.
Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. P. (2014). What's different about social media networks? A framework and research agenda. *MIS Quarterly, 38*(1), 275–304.
Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59–68. http://dx.doi.org/10.1016/j.bushor.2009.09.003.
Kasiviswanathan, S. P., Melville, P., Banerjee, A., & Sindhwani, V. (2011). Emerging topic detection using dictionary learning. In Proceedings of the 20th ACM international conference on Information and knowledge management − CIKM '11 745–754. New York, NY, USA: ACM. https://doi.org/10.1145/2063576.2063686.
Kefi, H., Mlaiki, A., & Kalika, M. (2015). Social Networking Continuance: When Habit leads to information overload. ECIS 2015 Proceedings, 1–13.
Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (4950 LNCS, pp. 154–175). https://doi.org/10.1007/978-3-540-70956-5_7.
Kepner, J., Gadepally, V., Michaleas, P., Schear, N., Varia, M., Yerukhimovich, A., & Cunningham, R. K. (2014). Computing on masked data: A high performance method for improving big data veracity. In 2014 IEEE High Performance Extreme Computing Conference, HPEC 2014. https://doi.org/10.1109/HPEC.2014.7040946.
Khare, P., Torres, P., & Heravi, B. R. (2015). What just happened? A framework for social event detection and contextualisation. In Proceedings of the Annual Hawaii International Conference on System Sciences (2015-March, pp. 1565-1574). https://doi.org/10.1109/HICSS.2015.190.
Kim, Y., Choi, K. S., & Natali, F. (2016). Extending the Network: The Influence of Offline Friendship on Twitter Network Full papers. *Proceedings of the Americas Conference on Information Systems*.
King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science, 331*(6018), 719–721. http://dx.doi.org/10.1126/science.1197872.
Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), http://dx.doi.org/10.1177/2053951714528481 [205395171452848].
Kleindienst, D., Pfleger, R., & Schoch, M. (2015). The business alignment of social media analytics. *ECIS 2015 Completed Research Papers*, 4801, Paper 103.
Kraft, T., Wang, D. X., Delawder, J., Dou, W., Yu, L., & Ribarsky, W. (2013). Less After-the-Fact: Investigative visual analysis of events from streaming twitter. In IEEE Symposium on Large Data Analysis and Visualization 2013, LDAV 2013 – Proceedings 95-103. https://doi.org/10.1109/LDAV.2013.6675163.
Kumar, S., & Rishi, R. (2015). Data collection and analytics strategies of social networking websites. In Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on 643-648. https://doi.org/10.1109/ICGCIoT.2015.7380543.
Lai, V., Rajashekar, C., & Rand, W. (2011). Comparing social tags to microblogs. In Proceedings – 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011 1380-1383. https://doi.org/10.1109/PASSAT/SocialCom.2011.52.
Li, H., Sakamoto, Y., & Chen, R. (2014). The Psychology Behind People's Decision To Forward Disaster- Related Tweets. In *PACIS 2014 Proceedings*.
Li, Z., & Huang, K.-W. (2014). The Monetary Value of Twitter Followers: Evidences from NBA Players. In Proceedings of the International Conference on Information Systems.
Lin, J. (2015). On Building Better Mousetraps and Understanding the Human Condition: Reflections on Big Data in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 33–47. http://dx.doi.org/10.1177/0002716215569174.
Liu, F. (2015). Retransmitting Messages Online in Evolving Disasters: A Scenario Simulation. *Proceedings of the International Conference on Information Systems*.
Liu, S.-H., Chou, C.-H., & Liao, H.-L. (2015). An exploratory study of product placement in social media. *Internet Research, 25*(2), 300–316. http://dx.doi.org/10.1108/IntR-12-2013-0267.
Liu, X., Tang, K., Hancock, J., Han, J., Song, M., Xu, R., . . . Pokorny, B. (2012). SocialCube: A Text Cube Framework for Analyzing Social Media Data. In 2012 International Conference on Social Informatics 252-259. https://doi.org/10.1109/SocialInformatics.2012.87.
Liu, X., & Huet, B. (2012). Social event discovery by topic inference. *International Workshop on Image Analysis for Multimedia Interactive Services,* 1–4. http://dx.doi.org/10.1109/WIAMIS.2012.6226752.
Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online, 24*, 4–15. http://dx.doi.org/10.7152/acro.v24i1.14671.
Lynn, T., Healy, P., Kilroy, S., Hunt, G., van der Werff, L., Venkatagiri, S., & Morrison, J. (2015). Towards a general research framework for social media research using big data. In 2015 IEEE International Professional Communication Conference (IPCC) 1–8. https://doi.org/10.1109/IPCC.2015.7235843.
Mahrt, M., & Scharkow, M. (2013). The Value of Big Data in Digital Media Research. *Journal of Broadcasting & Electronic Media, 57*(1), 20–33. http://dx.doi.org/10.1080/08838151.2012.761700.
Maynard, D., Bontcheva, K., & Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *Proceedings of the LREC 2012 Workshop @NLP can u tag #user_generated_content?!* 15-22. Retrieved from http://gate.ac.uk/sale/lrec2012/ugc-workshop/opinion-mining-extended.pdf.
McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review, 90*(October), 60–68.
Meth, S., Lee, K. Y., & Yang, S.-B. (2015). Factors Influencing Facebook Users' Political Participation: Investigating the Cambodian Case. In *PACIS 2015 Proceedings*.
Mirbabaie, Milad, Ehnis, Christian, Stieglitz, Stefan, & Bunker, Deborah (2014). Communication roles in public events: a case study on Twitter communications. In: Information systems and global assemblages: (re)configuring actors, artefacts, organizations: IFIP WG 8.2 Working Conference on Information Sysanizations, IS&O 2014, Auckland, New Zealand, December 11–12, 2014: proceedings/Doolin, Bill; Lamprou, Eleni; Mitev, Nathalie; McLeod, Laurie (Hrsg.) Heidelberg [u.a.] Springer, 207-218 ISBN: 978-3-662-45707-8 978-3-662-45708-5.
Mirbabaie, M., Tschampel, N., & Stieglitz, S. (2016). Geodaten in Social Media als Informationsquelle in Krisensituationen. In *Multikonferenz Wirtschaftsinformatik (MKWI) 2016* (Vol. 4, pp. 315-326).
Mirbabaie, M., & Zapatka, E. (2017). Sensemaking in Social Media Crisis Communication – A Case Study on the Brussels Bombings in 2016. *Proceedings of the Twenty-Fifth European Conference on Information Systems (ECIS)*.
Moalla, I., Nabli, A., Bouzguenda, L., & Hammami, M. (2017). Data warehouse design approaches from social media: Review and comparison. In Social Network Analysis and Mining (Vol. 7). https://doi.org/10.1007/s13278-017-0423-8.
Morstatter, F., Pfeffer, J., & Liu, H. (2014). When is it biased? In Proceedings of the 23rd International Conference on World Wide Web – WWW '14 Companion 555-556. https://doi.org/10.1145/2567948.2576952.
Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM,* 400–408. http://dx.doi.org/10.1007/978-3-319-05579-4_10.
Musaev, A., Wang, D., Shridhar, S., Lai, C. A., & Pu, C. (2015). Toward a Real-Time Service for Landslide Detection: Augmented Explicit Semantic Analysis and Clustering Composition Approaches. In Proceedings – 2015 IEEE International Conference on Web Services, ICWS 2015 511-518. https://doi.org/10.1109/ICWS.2015.74.
Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the European Parliament. *Electoral Studies*. http://dx.doi.org/10.1016/j.electstud.2016.04.014.
Oh, C., Hu, H.-f., & Yang, W. (2016). Social Media Information Diffusion and Economic Outcomes: Twitter Retweets and Box Office. In *PACIS 2016 Proceedings*.
Patel, A., Gheewala, H., & Nagla, L. (2014). Using social big media for customer analytics. In Proceedings of the 2014 Conference on IT in Business, Industry and Government: An International Conference by CSI on Big Data, CSIBIG 2014 1-6. https://doi.org/10.1109/CSIBIG.2014.7056974.
Payton, F. C., & Conley, C. (2014). *Fear or danger threat messaging: The dark side of social media*.
Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Proceedings of Human Language Technologies 2010* 181–189.
Pletikosa Cvijikj, I., Dubach Spiegler, E., & Michahelles, F. (2013). Evaluation framework for social media brand presence. *Social Network Analysis and Mining, 3*(4), 1325–1349. http://dx.doi.org/10.1007/s13278-013-0131-y.
Qian, S., Zhang, T., Xu, C., & Shao, J. (2016). Multi-Modal Event Topic Model for Social Event Analysis. *IEEE Transactions on Multimedia, 18*(2), 233–246. http://dx.doi.org/10.1109/TMM.2015.2510329.
Qin, Z., Cai, J., & Wangchen, H. Z. (2015). How rumors spread and stop over social media: A multi-layered communication model and empirical analysis. *Communications of the*

Association for Information Systems, 36, 369–391.
Rama Satish, K. V., & Kavya, N. P. (2014). Big data processing with harnessing hadoop – MapReduce for optimizing analytical workloads. In Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014 (pp. 49–54). https://doi.org/10.1109/IC3I.2014.7019818.
Rehman, N. U., Weiler, A., & Scholl, M. H. (2013). OLAPing social media: The case of Twitter. *International Conference on Advances in Social Networks Analysis and Mining*, 1139–1146. http://dx.doi.org/10.1145/2492517.2500273.
Reuter, T., & Cimiano, P. (2012). Event-based classification of social media streams. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval – ICMR '12. New York, NY, USA: ACM. https://doi.org/10.1145/2324796.2324824.
Ritter, A., Etzioni, O., & Clark, S. (2012). Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '12 (p. 1104). https://doi.org/10.1145/2339530.2339704.
Ruiz, M. C., Calleja, J., & Cazorla, D. (2015). Petri Nets Formalization of Map/Reduce Paradigm to Optimise the Performance-Cost Tradeoff. In 2015 IEEE Trustcom/BigDataSE/ISPA (Vol. 3, pp. 92–99).
Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064. http://dx.doi.org/10.1126/science.346.6213.1063.
Saha, B., & Srivastava, D. (2014). Data quality: The other face of Big Data. In Proceedings – International Conference on Data Engineering (pp. 1294–1297). https://doi.org/10.1109/ICDE.2014.6816764.
Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social Media Analyses for Social Measurement. *Public Opinion Quarterly, 80*(1), 180–211. http://dx.doi.org/10.1093/poq/nfv048.
Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big Data, Digital Media, and Computational Social Science: Possibilities and Perils. *The Annals of the American Academy of Political and Social Science, 659*(1), 6–13. http://dx.doi.org/10.1177/0002716215572084.
Shen, Y., Hock Chuan, C., & Cheng, S. H. (2016). The Medium Matters: Effects on What Consumers Talk about Regarding Movie Trailers. In Proceedings of the International Conference on Information Systems.
Simmonds, R. M., Watson, P., Halliday, J., & Missier, P. (2014). A Platform for Analysing Stream and Historic Data with Efficient and Scalable Design Patterns. In 2014 IEEE World Congress on Services (SERVICES) (pp. 174–181). https://doi.org/10.1109/SERVICES.2014.40.
Song, M., & Kim, M. C. (2013). RT2M: Real-time twitter trend mining system. In Proceedings – 2013 International Conference on Social Intelligence and Technology, SOCIETY 2013 (pp. 64–71). https://doi.org/10.1109/SOCIETY.2013.19.
Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining, 3*(4), 1277–1291. http://dx.doi.org/10.1007/s13278-012-0079-3.
Stieglitz, S., Bunker, D., Mirbabaie, M., & Ehnis, C. (2017a). Sense-Making in Social Media During Extreme Events. *Journal of Contingencies and Crisis Management (JCCM)*. http://dx.doi.org/10.1111/1468-5973.12193.
Stieglitz, S., Brachten, F., Ross, B., & Jung, A. (2017b). Do Social Bots Dream of Electric Sheep? A Categorisation of Social Media Bot Accounts. *Proceedings of the 28th Australasian Conference on Information Systems (ACIS)*.
Stieglitz, S., Mirbabaie, M., Schwenner, L., Marx, J., Lehr, J., & Brünker, F. (2017c). Sensemaking and Communication Roles in Social Media Crisis Communication. *Proceedings of the 13th International Conference on Wirtschaftsinformatik (WI)*.
Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics – An Interdisciplinary Approach and Its Implications for Information Systems. *Business & Information Systems Engineering, 6*(2), 89–96. http://dx.doi.org/10.1007/s11576-014-0407-5.
Susarla, A., Oh, J.-H., & Tan, Y. (2012). Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube. *Information Systems Research, 23*(1), 23–41. http://dx.doi.org/10.1287/isre.1100.0339.
Tang, J., Hu, X., Gao, H., & Liu, H. (2012). Unsupervised feature selection for linked social media data. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '12 (pp. 904–912). https://doi.org/10.1145/2339530.2339673.
Tinati, R., Phillipe, O., Pope, C., Carr, L., & Halford, S. (2014). Challenging Social Media Analytics: Web Science Perspectives. In Proceedings of the 2014 ACM Conference on Web Science (pp. 177–181). New York, NY, USA: ACM. https://doi.org/10.1145/2615569.2615690.
Tsou, M.-H., Jung, C.-T., Allen, C., Yang, J.-A., Gawron, J.-M., Spitzberg, B. H., & Han, S. (2015). Social Media Analytics and Research Test-bed (SMART Dashboard). In: SMSociety '15, Proceedings of the 2015 International Conference on Social Media & Society (2:1–2:7). New York, NY, USA: ACM. https://doi.org/10.1145/2789187.2789196.
Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 505–514.
Vaccari, C., Valeriani, A., Barber, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. (2013). Social media and political communication. A survey of Twitter users during the 2013 Italian general election. Rivista italiana di scienza politica, 381–410. https://doi.org/10.1426/75245.
Valkanas, G., Katakis, I., Gunopulos, D., & Stefanidis, A. (2014). Mining Twitter Data with Resource Constraints. In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) (pp. 157–164). https://doi.org/10.1109/WI-IAT.2014.29.
van Gorp, A., Pogrebnyakov, N., & Maldonado, E. (2015). Just Keep Tweeting: Emergency Responder' Social Media Use Before and During Emergencies. In Proceedings of the European Conference on Information Systems. https://doi.org/10.18151/7217512.
van Osch, W., & Coursaris, C. K. (2013). Organizational Social Media: A Comprehensive Framework and Research Agenda. In 2013 46th Hawaii International Conference on System Sciences (HICSS) (pp. 700–707). https://doi.org/10.1109/HICSS.2013.439.
Vavliakis, K. N., Symeonidis, A. L., & Mitkas, P. A. (2013). Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering, 88*, 1–24. http://dx.doi.org/10.1016/j.datak.2013.08.006.
Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. MIS Quarterly, 37(1), 21–54.
Venkatesh, V., Brown, S., & Sullivan, Y. (2016). Guidelines for Conducting Mixed-methods Research: An Extension and Illustration. *Journal of the Association for Information Systems, 17*(7), 435–495.
vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. ECIS 2009 Proceedings.
vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, 37(1).
Wang, J. H., & Chen, K. T. (2015). Towards an efficient platform for social big data analytics. In Wireless and Optical Communication Conference (WOCC), 2015 24th (pp. 175–179). https://doi.org/10.1109/WOCC.2015.7346200.
Wang, N., Ding, N., & Yang, D. (2014). Containment of Misinformation Propagation in Online Social Networks With Given Deadline. In *PACIS 2014 Proceedings*.
Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly, 26*(2), xiii–xxiii.
Wegrzyn-Wolska, K., Bougueroua, L., & Dziczkowski, G. (2011). Social media analysis for e-health and medical purposes. In Proceedings of the 2011 International Conference on Computational Aspects of Social Networks, CASoN'11 (pp. 278–283). https://doi.org/10.1109/CASON.2011.6085958.
Weiler, A., Grossniklaus, M., & Scholl, M. H. (2016). Situation monitoring of urban areas using social media data streams. *Information Systems, 57*, 129–141. http://dx.doi.org/10.1016/j.is.2015.09.004.
Wendling, C., Radisch, J., & Jacobzone, S. (2013). The Use of Social Media in Risk and Crisis Communication. OECD Working Papers on Public Governance No. 24. (24), 1–42. https://doi.org/10.1787/5k3v01fskp9s-en.
Weng, J., Yao, Y., Leonardi, E., Lee, F. (2011). Event Detection in Twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (pp. 401–408). https://doi.org/10.1109/ICTAI.2007.23.
Yang, C. C., & Ng, T. D. (2011). Analyzing and visualizing web opinion development and social interactions with density-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans, 41*(6), 1144–1155. http://dx.doi.org/10.1109/TSMCA.2011.2113334.
Yin, S., & Kaynak, O. (2015). Big Data for Modern Industry: Challenges and Trends. Proceedings of the IEEE, 103(2), 143–146. https://doi.org/10.1109/JPROC.2015.2388958.
Yu, G., & Zou, D. (2015). Which User-Generated Content Should Be Appreciated More? – A Study on UGC Features, Consumers' Behavioral Intentions and Social Media Engagement. ECIS 2015 Proceedings. Paper 211.
Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems, 25*(6), 13–16. http://dx.doi.org/10.1109/MIS.2010.151.
Zhang, H., Sun, Z., Liu, Z., Xu, C., Wang, L. (2015). Dart: A Geographic Information System on Hadoop. In Proceedings – 2015 IEEE 8th International Conference on Cloud Computing, CLOUD 2015 90-97. https://doi.org/10.1109/CLOUD.2015.22.
Zhang, S., Zhao, L., Lu, Y., & Yang, J. (2016). Do you get tired of socializing? An empirical explanation of discontinuous usage behaviour in social network services. Information & Management. Advance online publication. https://doi.org/10.1016/j.im.2016.03.006.
Zhang, Z., & Zhang, Y. (2016). How Do Explicitly Expressed Emotions Influence Interpersonal Communication and Information Dissemination? a Field Study of Emoji ' S Effects on Commenting and Retweeting on a Microblog Platform. In *PACIS 2016 Proceedings*.
Zhao, Z., Feng, Z., Zhang, Y., Ning, L., Fan, J., & Feng, S. (2015). Collecting Managing and Analyzing Social Networking Data Effectively. In 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) 1642–1646. https://doi.org/10.1109/FSKD.2015.7382191.