Completeness in Twitter datasets

A critical review on Twitter research methodologies

JOOST POTTING

A thesis submitted in partial fulfilment of the

requirements for the degree of

New Media and Digital Culture

Utrecht University

j.potting@students.uu.nl // joost.potting@xx.com

5595134

Amsterdam 17 June 2016

First reader Mirko Tobias Schäfer

Second reader Stefan Werning

Summary

This thesis addresses how the limitations of Twitter research influences the datasets, and this is discussed in the literature. Twitter is used for a range of topics, from tracking countries online to predicting epidemics. This paper researches how the limitations of using Twitter as a data source is discussed, and how completeness influences the research. By analyzing the methodology used in a study of the Dutch Twittersphere as a case study, several limitations are revealed. The current practices of conducting Twitter research are analyzed, and an overview of the current discussion about limitations is presented. This is followed by an introduction of completeness in Twitter research. Completeness is discussed on three levels, the literal definition, disclosure of methodology, and addressing limitations in research. The influence of these three levels on the replicability of research is discussed, and consequently a proposal for a model is presented.

Foreword

Firstly, I would like to thank Mirko Tobias Schäfer for his guidance and help over the past months and Stefan Werning for serving as a second reader for my thesis. I also want to express my gratitude to my direct colleagues, who supported me, provided me with recommendations, and discussed the complex matter with me. Finally, writing this thesis would not have been possible without the unconditional belief and help of my family and friends. Even though they found most topics rather abstract and complex, their support was of great importance in my successfully writing this thesis.

Table of Contents

1. Introduction

Since the inception of Twitter in 2006 (Twitter, 2016), researchers have examined the service, the users, and the messages sent. As Zimmer and Proferes (2014) noted, between 2007 and 2012, 382 papers focused on Twitter, spanning 17 different topics (p. 253). Williams, Terras, and Warwick (2013) classified 575 research papers that focused on Twitter, based on their abstract. They discern four different analysis methods that span 13 categories. Both studies demonstrate that Twitter is a source of information for researchers interested in a variety of topics. Within the humanities, Twitter research can be placed in big data research, the act of researching large datasets (Manovich, 2011). As there is a clear indication that Twitter is a focus of study, it is important that the used methodologies and resulting limitations be addressed. By analyzing a case study, this thesis analyzes and reflects on the use of Twitter data in the Digital Methods. With an increasing use of datasets in the humanities (see Borgman, 2009; boyd & Crawford, 2012; Burdick, Drucker, Lunenfeld, Presner, & Schnapp, 2012; Manovich, 2011), a critical look at the used methodologies and limitations of data handling is important.

In the foreword of his book, the communications scholar Steve Jones introduced the field of internet research (1999). He introduced and asked the other authors of the book to research internet in a social science perspective. The chapters focus on the discussion of internet research methods as it had been practiced. Ten years after the publication of Jones' book, Richard Rogers proposes that internet research can benefit from genuine Digital Methods (2009, p. 5). This transition consists of developing methods to research digital objects of the internet instead of online culture or digitized data. Rogers specifically focuses on the limitations of the current methods.

As demonstrated in the following chapter, how data is handled and which methods are used differ among research fields. This paper attempts to answer the following research questions:

> How are the limitations of using Twitter as a data source discussed in the literature?
>
> How is completeness of Twitter research discussed and how does it influence the research?

The methods, reported methodologies, and the limitations of using Twitter as a data source are analyzed. Twitter is used in many ways to research phenomena, but it also has its limitations. Different researchers discuss these limitations. This paper is structured as follows: First, the scientific positioning is discussed. This is followed by a case study that is introduced and discussed. With the case study in mind, the current practices of conducting research with Twitter data are analyzed. Subsequently, the limitations that Twitter researchers encounter are analyzed. These limitations are complemented by a chapter that introduces completeness as an influence on the replicability of Twitter research. This is followed by a proposal for a model that focuses on this aspect. Lastly, a conclusion is formulated, as well as a reflection on this thesis.

2. Scientific Positioning

2.1 Media Studies and Digital Methods

The media scholar Lisa Gitelman, in co-authorship with Virginia Jackson, states in the introduction to the book *Raw Data is an Oxymoron* that humanity students and scholars are alienated from data (Gitelman & Jackson, 2013, p. 3). However, they propose that those working within the humanities need to make informed decisions regarding data. They furthermore highlight the possibilities for humanists with regard to big data, especially the large scale and possession of digital data. Burdick et al. (2012) introduce digital media as a new source of interest within the digital humanities[1], which requires close reading[2]. Rogers (2013) introduces two categories of data in digital media, natively digital data, and digitized data. Digitized data is digitally constructed data and data that is transformed into digital data, such as a document scan. As defined by Manovich (2011), digital humanities traditionally used digitized data, such as newspapers, books, and pictures. However, as Rogers argues in his book on Digital Methods (2009), natively digital data, such as Twitter data, needs to be handled differently. Four year later, Rogers (2013) argues that it is important that research regarding the online web focuses on the study of natively digital data and the methods used to analyze it. He furthermore proposes the following:

> Follow the methods of the medium as they evolve, learn from how dominant devices treat natively digital objects, and think along with those object treatments and devices so as to recombine or build on top of them. Strive to repurpose the methods of the medium for research that is not primarily or solely about online culture (Rogers, 2013, p. 5).

Rogers hereby encourages researchers to experiment with the methods and to build upon established methods.

The same observation is made by Burdick et al. (2012). They describe Digital Methods as developing, but also stabilizing. They further mention that experimental processes are necessary, and the standardization and normalization of methodologies and practices should not be rushed (p. 21). An example of this phenomenon is demonstrated in the book *Twitter and Society,* where research based around Twitter is discussed (Weller, Bruns, Burgess, Mahrt, & Puschmann, 2014). Some of the discussed research projects use the same methods, but most are in an experimental phase. This furthermore validates the relevance to discuss and reflect upon the methods in Twitter research. The editors of the book explain that research based on Twitter not only provides insights to the platform and its users, but also on society as a whole (p. xxxi).

---

[1] See both Burdick et al. (2012) and Berry, (2012) for an overview of Digital Humanities.
[2] Close reading refers to analyzing the specific features of any individual text, contrary to distant reading, which examines larger patterns from a corpus of text (Burdick et al., 2012, p. 39).

2.2 Twitter Research

As argued by Vis (2013), research based around Twitter is a field within big data research, which is presented in the Digital Methods. In the case of Twitter research, it is common to gather large quantities of data that fit the definition of big data as described in Chapter 2.3. Twitter has been used in a range of different fields, ranging from communities on Twitter (Bruns, Burgess, & Highfield, 2014), medical epidemics (Collier, Son, & Nguyen, 2010), happiness (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011), politics (Paßmann, Boeschoten, & Schäfer, 2014), and the platform itself (Bruns & Stieglitz, 2013; Kwak, Lee, Park, & Moon, 2010; Miller, Ginnis, Stobart, Krasodomski-Jones, & Clemence, 2015). Zimmer and Proferes (2014) provide an overview of Twitter research, and introduce and discuss several themes and practices. They analyzed 382 studies within their scope, number of tweets, collection methods, and ethical considerations. Zimmer and Proferes (2014) categorized the articles according to seventeen topics, which indicates the different interests of researchers. One year prior, Williams et al. (2013) examined the summaries of Twitter research. Based on 575 articles, the authors identified the domain, methodologies, aspects of research (focus on messages, users, technology, or concept), and the data used. They concluded that most of the summaries of papers focus on the messages sent and the details of the users. Moreover, the different categories of interest did not have a specific methodology preference, except for one[3].

In the introduction to *Twitter and Society*, Rogers (2014) describes three phases of Twitter research. He distinguishes three periods of Twitter and resulting research: past, present, and future. Past research defined Twitter as a social network used for "pointless babble". Researchers examined tweet content and analyzed the interactions between users. Current research, at least at the time of publishing his article, examined Twitter as a news medium. More often, Twitter was ahead of traditional media in reporting evolving news, ranging from attacks to births or worldwide events. Researchers examined the way in which information spreads via Twitter, the disadvantages, and the (possible) advantages of this phenomenon. The future of Twitter research, as described by Rogers, revolves around archived data. Twitter is more often approached as a dataset with archived information. As a result, researchers have to be able to use technical infrastructure for gathering, storing, and analyzing tweets (p. xxi). The future of Twitter research, as argued by Rogers (2014), revolves around studying the Application Programming Interface (API), databases, and technological challenges. However, as described earlier, the methods employed are still evolving and are relevant for study, as purposed by this paper.

2.3 Data Turn

Archived data, which Rogers argues is the future of Twitter research, is essentially datasets that can be analyzed. Researchers in different fields, as seen in the previous chapter, have already used such datasets. The studies used datasets of different sizes in terms of messages, but also other relevant data (e.g. time of tweet sent, location, reply to etc.). In their article on big data, danah

[3] Williams et al. (2013) identified that papers concerning 'Libraries', including archives and repositories, most often used examination as a methodology.

boyd and Kate Crawford describe a critical question about this concept, which also can be used for Twitter research (boyd & Crawford, 2012). The authors note that there is no size limit, either upper or lower, for datasets to be considered big data. In addition, they remark that what is or is not considered big data could change with time (p663). boyd and Crawford provide several examples of how big data and social media platforms are influenced[4].

With the emergence of social media, and their large datasets, Manovich (2011) notes that opportunities emerged for researchers to gain insights into the behavior of a large number of people. A benefit is that although they know their data can be used for commercial purposes, they probably do not realize their data can be subject to study. This data has many visible, but also invisible attributes, and is often described as big data (Bruns & Stieglitz, 2013). As a result, those working with social media data have to consider the challenges that working with big data brings. Laney (2001) described three aspects of big data: volume, velocity, and variety. Volume refers to the sheer size of datasets, velocity to the speed of creating new data, while variety refers to the different data forms. In addition to the technical description of big data, boyd and Crawford (2012) propose a three-part definition, which encompasses a cultural, technological, and scholarly phenomenon. They find that big data is less about the data itself and more about the "capacity to search, aggregate, and cross-reference large datasets" (p. 663). Thus, they define big data as the interaction between technology, analysis, and mythology.

In her article on big data, Vis (2013) complements the definition of analysis made by boyd and Crawford (2012). Firstly, she introduces validity, which refers to how a sample is constructed and what can be deduced from this sample as a result. She highlights the methods researchers most commonly use to access Twitter[5] and how their sample relates to the general population. Secondly, Vis introduces venture, whereby she addresses how researchers handle the curiosity about the data at hand, the exploration of this data, and its interpretation. Researchers aim to make a point when discussing their data. The final definition Vis introduces is visibility, which refers to the methods of creating and dealing with the data, which generally remain invisible to others. It furthermore refers to visualizations of data: what it actually reveals and what it hides. Vis notes that it is important for researchers to clarify the specific visualizations and methods used.

boyd and Crawford (2012) also describe how critical data handling is. With the automation of collecting and analyzing data, it is important to distinguish which methods are used and how they are regulated. They explain that these methods and regulations will shape the future of big data analysis. In addition to this concern, boyd and Crawford (2012) discuss six issues surrounding big data. Their sixth issue regards the limited access to big data. Access to data is limited to researchers with funds to buy data or those who are part of a company that has such funds, and those who do not have access to funds. This results in researchers who can analyze a wealth of

---

[4] For example: data cleaning, p. 667; connections between messages and accounts, p. 670, friends and followers, p.671; ethical points, p. 672; accessibility of data, p. 673

[5] Accessing Twitter via the Application Programming Interface (API) is discussed in Chapter 5.3

information that others cannot verify or reproduce. In addition to this divide, boyd and Crawford (2012) make a distinction between researchers with technological skills and expertise to gather and analyze datasets and those who do not. As a result, they concluded that a digital divide between those with access and those without access is emerging.

The debate on big data encompasses a wide range of fields. Ekbia et al. (2015) demonstrate that because of the absence of a clear definition of big data, different fields use different perspectives and methods. In addition, regarding methodology, the researchers highlight several issues with the current state of research. They provide an overview of remedies for these issues, but remark that data selection is still a concern. As summarized by the researchers, the selection or omission of specific data can result in different statistical outcomes. This could enable researchers to use only data that supports their hypothesis. As an example, they discuss the sampling of Twitter data. Researchers can conduct this sampling deliberately, but it is more likely that Twitter limits the possibilities for researchers to access the data. This selection bias, as well as other limitations, is discussed in the literature overview in chapter 5.

3. Methodology of Research & Case Study

The case study at hand focuses on previous research by highlighting limitations by means of a comparative analysis. The aim of this mixed methods design is to highlight limitations in a case study and compare these to other research projects. The comparative analysis consists of reflecting on the identified limitations in the case study and comparing them to current Twitter research practices. Consequently, the practices will be reflected upon by examining the discussion in the literature. Firstly, a research project by students of the Utrecht Data School is analyzed. This is followed by a critical reflection on the current Twitter research practices. In chapters 4 to 7, several research papers are discussed and analyzed with regard to the methodology used. Consequently, chapter 7 proposes a model in which the current research methodologies are placed in the perspective of completeness.

3.1 Methodology of Case Study

The Utrecht Data School is a research project based at the University of Utrecht that performs data analyses in different forms[6]. As part of a project commissioned by the University of Applied Sciences Utrecht, an attempt was made to download all Dutch tweets that were sent in one week. This case study describes how this data was downloaded, verified, and improved.

The project aimed to map the Dutch Twittersphere, and especially the local media ecosystems within the Twittersphere[7]. The original research team received their dataset from a third party partner. This partner used a word list with 37,633 Dutch terms, which was used to download the information. In an attempt to validate the dataset, the project team analyzed 24 Twitter accounts manually[8]. The researchers estimated that 58.5% of all Dutch tweets on Twitter had been downloaded. The methodology can be positioned within Digital Methods. As mentioned by Burdick et al. (2012), Digital Methods are stabilizing but are still in an experimental state. The methodology used in this case study is in line with other Twitter research, but is still experimental. However, the aim for the used method was to be as accurate as possible within the imposed limitations. In chapter 4, different papers on Twitter research are discussed, as well as the methods used.

3.1.1 Calculating the Completeness

With the aim of finding and downloading a higher percentage of all Dutch tweets, a new list of words was proposed. In addition, the inclusion percentage was recalculated in a more sophisticated manner using an ego-based method. This method consists of selecting a sample of random users from the complete dataset, downloading all their tweets, and comparing the tweets to the complete dataset. Contrary to Krijger et al. (2016), the users were selected randomly. In his paper that discusses how to determine the sample size, Israel (1992) argued that for a total

---

[6] See www.dataschool.nl/home-2/ for more information on the projects.
[7] The report by Krijger et al. (2016) is available upon request via the author or www.dataschool.nl/home-2
[8] The method consisted of handpicking Twitter accounts with "ij" in the name. The authors claim this is typically Dutch. Consequently, all tweets from these accounts were downloaded and compared to the complete dataset.

population of over 100,000 individuals, a sample size of 400 is sufficient. Unfortunately, Twitter does not release information on how many users each country or language has, so an estimation has to be made. The original dataset consisted of 563,050 unique users, which provides an indication of the total Dutch Twitter population. For this reason, a sample of 400 users were randomly selected (See Appendix II: Randomly Selecting Users). From these users, all available tweets were downloaded via a third party web service[9] (See Appendix III). This resulted in 339 users with more than one tweet sent in the investigated week. After examining the data in detail, it was discovered that 38 users had not sent any tweets. A manual investigation revealed that these users did send more than 3,200 tweets between the end of the specific week and the current research and as a result the tweets from the specific week were not downloaded[10]. This number of 3,200 tweets is set by Twitter as a maximum limit of returned tweets. To compensate for missing these tweets, the average tweets per week was calculated and added to the analysis. Of these 38 users, 16 were either in another language or identified as a spam account[11] and were filtered out. Another 37 users were omitted from the automatically generated dataset. After manually analyzing each account, it was revealed that twenty accounts were either set to private, deleted, or banned from the site. Furthermore, two users did not tweet in Dutch and one account was created in May 2016, six months after the original study. Especially the last account demonstrates that Twitter is constantly changing, and repeating a previous research is difficult. After adding the manual counts of the tweets, 361 accounts remained. By comparing all tweets by the selected users to the complete dataset, it was calculated that 59.34% of the tweets were present in the original dataset. This result demonstrates that even though Krijger et al. (2016) made a fairly accurate estimation, it is importance to make statistically sound conclusions.

3.1.2 Improving the Word List

Improving the word list was part of a previous research project. This consisted of adding the most frequent words already present in the downloaded dataset, the most common Dutch verbs and city names, as well as the most frequently used words in Dutch corpora[12,13]. The process of adding new terms to the word list did reveal limitations, both from the commercial partner side as well as Twitter. Understandably, the partner wanted to keep the list as short as possible to limit the technological impact a long list has on their servers. In addition, it clarified that they search by key word, one at a time. Twitter and their APIs impose this limitation.

---

[9] At the time of writing, www.twitonomy.com allows users to purchase a subscription that allows them to analyze any Twitter account. This includes downloading all the available tweets.
[10] These users did send more tweets than 3,200 between the end of the investigated week (19 – 25 November 2015) and the current research (June 2016). As a result, the Twitter API does not return tweets send in the investigated week.
[11] A spam account sends a high volume of tweets per day, mostly consisting of current trending topics with a URL to a specific site.
[12] This research project is available upon request.
[13] These are made available by The Institute for Dutch Lexicology and consist of the top 5000 words of nine individual corpora.

3.2 Analysis of Methodology

The methodology, as described, is experimental in a sense. There is no standardized method available, only descriptions on how others have attempted to achieve the same or similar results. This influences the results, but it furthermore indicates the limitations. These limitations are further discussed and analyzed. Firstly, how the word list was expanded is analyzed, followed by the process of calculating the completeness.

3.2.1 Word List

The main limitation of the word list is the way in which Twitter allows researchers to download data. Researchers must provide a list of words or hashtags they want to track and download. This influences the completeness of the dataset, as discussed in chapter 6. The main problem with adding words is the selection of the words, and especially those omitted. With the correct motivation, the decisions can be debated. This demonstrates the importance of addressing these limitations, which is likewise addressed in this thesis.

In addition, dependence on the third party introduces another limitation. As the exact process of constructing the dataset is unknown, a so-called black box is introduced. A black box refers to the process in which one inputs data, the black box does something, and a result is presented. The "something" this program does, remains unknown. Chapter 5.2 further highlights and reflects on this aspect.

3.2.2 Calculating Completeness

The methodology for calculating the completeness reveals three major limitations. The first is the selection bias and which users to select. The original dataset consisted of eight files, one for each day. Fifty users were randomly selected from each file. Thus, this selection has its sample bias, as active users have a greater chance to be included in the selection. However, as active users are more present, if they do not appear in the selection, they influence the completeness of the dataset in a greater measure.

The second limitation determined is the use of a black box, in this case a web service. The used commercial service does not provide insight into their source code or methodology. Even though this is a valid black box, a close examination of the downloaded tweets revealed that no algorithm or filter was applied to the tweets. If technological expertise were available, it would have been possible to develop a script or tool to perform the same task. It highlights another aspect of the two-cultures problem as postulated by Rieder and Röhle (2012) and boyd and Crawford (2012). Rieder and Röhle (2012) refer to 'code literacy', the ability of researchers to comprehend and potentially develop code themselves. In this case study, even though the development of the code was not possible, the underlying methodology is understood. boyd and Crawford (2012) discuss the division between the researchers who are able to buy data, and those who are not. A one-month subscription does not cost much[14], but introduces the limitation of those able to afford a subscription and those who are not.

[14] At the time of writing a one-month subscription costs $20.

The problem of not being able to download the tweets of those users with more than 3,200 tweets introduces the third limitation. Should researchers want to replicate a Twitter study from the start (e.g. downloading all the tweets), there is a time constraint when analyzing popular users. For this case study, there was a time difference of six months between the initial dataset and downloading individual users' tweets. The deleted accounts indicate that replicating a study is difficult, as the environment is constantly changing. As Bruns remarks: "[…] a scholarly publishing industry in which journal articles and book chapters can sometimes take more than two years from submission to publication" (2013, para. 3.2). It is conceivable that it is virtually impossible for researchers to repeat some studies after a certain amount of time. Should researchers want to determine the number of tweets, they could follow a similar methodology as used in this case study, by calculating an average. However, this is only an estimation and does not allow analysis of message content.

3.3 Conclusion

This case study demonstrates that repeating (parts) of a previous research presents certain limitations. This is by either the used methodology or the limitations imposed by Twitter regarding the use of their API. The study also introduces an experimental method to calculate the completeness of a dataset. Future research must further develop the method and evaluate the validity of the method. By including the extensive methodology as an appendix (Appendix II and III), other researchers can repeat the process more easily and evaluate the steps taken. It furthermore complies to the concept of validity as introduced by Vis (2013). It refers to how a dataset is constructed and what potentially can be deduced from this sample.

It is possible that other researchers have developed a method to assess the completeness of their Twitter dataset[15]. However, as there is no standardized method available, each research project has to develop its own method. This does not only influence the validity of the different methods used, it is also more time consuming. By including the methodology used in this case study, researchers can potentially utilize the same method. It also reveals the decisions concerning the limitations.

When designing a dataset researchers have to consider several limitations. The case study highlighted one of these in particular, completeness. To analyze all Dutch tweets over a one-week period, it is necessary to determine the completeness. However, other limitations also influence the methodology. As Rosenberg (2013) concludes in his chapter on the historic definition of data:

> […] we should all be ready to engage with quantitative humanities approaches in a strong, critical fashion. Among other things, as humanists, we need to pay much better attention to the epistemological implications of search, an entirely new and already dominant form of inquiry, a form with its own rules, and with its own notable blind spots both in design and use (Rosenberg, 2013, p. 35).

[15] However, extensive search did not find such a method.

The influence blind spots have on Twitter research has been studied by several researchers. These limitations especially influence the research methodology. Chapter 5 discusses these limitations. Most of the discussed papers highlight the implications of the used tools and data access on Twitter research. In addition, the completeness of the dataset is discussed, as well as the replicability. The observations and results of the case study are highlighted in chapter 5.

4. Methods for Conducting Twitter Research

As mentioned in the introduction, Twitter has been examined from different aspects. These aspects can broadly be placed in one of three categories, each of which uses different methodologies. The first, and one of the more common, is examining the messages Twitter users sent. This typically revolves around a specific topic, theme, or activity based on hashtags or keywords. The second category is analyzing the users. Again, this can revolve around a specific topic or theme, but also countries or activities. The last and broadest, in terms of subjects, focuses on Twitter as a medium. This field examines the used technologies such as the API.

As will be demonstrated, using hashtags or keywords as a basis for Twitter research has been used several times. Unfortunately, it is not common to include the criteria for selecting the words. Ausserhofer and Maireder (2013) attempted to map the Austrian political Twittersphere. They addressed that not all tweets contain hashtags regarding a certain topic. By using a list of topic keywords, the researchers identified 1,657 accounts discussing those topics. However, as the research encountered technical limits of the Twitter API, they were

> forced [...] to narrow this user base, and to that end, we decided only to include accounts that (a) had more than 100 followers, (b) had 'hit' at least two political keywords or hashtags with their tweets and (c) were listed at least once by others (Ausserhofer & Maireder, 2013, p. 297).

They did not explain what was meant by "listed" or how this influenced their research. Eventually, they analyzed 374 Austrian users by downloading all their tweets. The authors mention in their conclusion that by using a "multiphase, user-centred [sic] approach allowed us to both avoid the problems inherent to many hashtag-based studies and to reliably identify important topics and political actors on Twitter" (Ausserhofer & Maireder, 2013, p. 308). Unfortunately, they do not include the list or number of keywords they tracked, how their self-designed tracker for downloading all tweets operated, or how they handled the small sample of users. They did distribute their tracker as an open source, revealing one step of their methodology, which makes it easier for researchers to replicate their study.

Some researchers include the keywords used in their research, but do not include the reasoning for selection. Collier et al. (2010) used Twitter to track influenza outbreaks. They downloaded 225,000 tweets by selecting tweets with one of seven keywords related to influenza. Although these keywords, or hashtags, do conform to the topic (e.g. *flu, influenza, swine flu, etc.*), the users tweeting about their influenza but without the specific hashtags would have been omitted. In addition, the author aimed to provide insight into users avoiding locations due to the risk of contamination. However, they could only conclude that the study demonstrated a high correlation between social media posts and diagnostic data. They also concluded that this method of tracking posts could be a method for a low-cost network regarding illness. By not providing motivation as to why they chose the keywords and how they would handle messages without the specific keywords, which were thus overlooked in the dataset, such a conclusion should be complemented with the observed limitations regarding the keywords.

It is furthermore possible that researchers do not want to investigate a certain topic, but a sample of messages on a random subject. With their aim to determine happiness in Twitter messages, Dodds et al. (2011) analyzed 4.9 billion tweets. Unfortunately, they do not include a methodology in their research. They only note when, how many tweets, and how many unique users were downloaded. Analysis of their description of how Twitter handles data reveals that they may have used one of three APIs that randomly selects 1% of all tweets. Dodds et al. (2011) do include limitations to their methods and address the issue of representativeness of their dataset. They note that the downloaded tweets were randomly selected by Twitter and did not include all tweets by all users. On the contrary, they note that collecting all tweets results in "a non-uniform subsampling of all utterances made by a non-representative subpopulation of all people" (p. 2). In other words, if researchers are able to analyze all tweets in their selected period, they still only analyze those users and messages sent on one platform, as discussed in chapter 5.1, and which is an aspect of selection bias.

Analyzing Twitter users is the second largest category within Twitter research[16]. Researchers attempting to analyze a specific group of users should always aim at a dataset that is as complete as possible. Bruns et al. (2014) used a mixed methodology in their research into all Australian Twitter users. They started by tracking down all users who were discussing typical Australian topics and subsequently used a snowball method to track down the followers and followees of those users. Their main method of filtering Twitter users from Australia is by only including those who have set an Australian time zone in their profile. Another assumption made by the researchers is that Australian Twitter users are predominantly connected to other Australian users. As a result, users who did not specify their time zone, were not connected to other Australian users, and did not participate in the selected topics were not included in the dataset. The researchers did not address this selection bias or discuss how it could be addressed. The authors did specify how they tracked down the users and applied the selection criteria. However, they did not include an extended methodology, which makes it difficult for other researchers to replicate their study and findings. In addition, Bruns et al. (2014) address the limitations of the Twitter API, but do not specify the API used. As discussed, it is possible that Twitter applied an algorithm to the results. Interestingly enough, Bruns et al. (2014) do address their concerns with, and the limitations of, hashtag-based research.

Kwak et al. (2010) started their research three years after the launch of Twitter (Twitter, 2016). They downloaded all users following a well-known celebrity, which was followed by downloading their users using a snowball method. In addition to these users, they tracked Twitter users tweeting about popular subjects but who were not connected to the celebrity. Kwak et al. (2010) reached interesting conclusions, such as that a retweeted tweet on average reaches 1,000 users. However, they did not include the same limitation as Dodds et al. (2011) regarding the complete dataset.

[16] See Williams et al., table 4, p. 12 (2013)

Miller et al. (2015) describe how they developed tools to collect, store, and analyze Twitter datasets with regard to the reliability of the dataset. In their report they do not describe the details of those tools, but instead refer to Wibberley, Reffin, and Weir (2014). To ensure they downloaded all relevant data, they used an over-expansive set of keywords (p. 17) in an attempt to limit the risk of overlooking valuable and relevant data. They address the fact that it is not possible to determine exactly what data was omitted. Miller et al. (2015) report on the relative ease of filtering irrelevant data by using computational methods. Lastly, they address the sampling limitation by comparing their (Twitter) data to offline data (e.g. questionnaires). By using self-developed tools, the researchers limit that black box for themselves. However, by not releasing the tool or source code, it is not possible to evaluate the used tools objectively.

Lastly, this chapter examines the research discussed in the book *Twitter and Society* (Weller et al., 2014). As discussed, the editors of the book remark how methodologies in Twitter research continue to develop in different research areas. This is repeated in the fifth chapter of the book *Data Collection on Twitter,* in which the authors discuss the different APIs Twitter provides and the tools for collecting the data (Gaffney & Puschmann, 2014). The limitations of the APIs and tools are discussed in the concluding paragraph of the fifth chapter. In particular, Gaffney and Puschmann (2014) remark that for the best result, researchers using Twitter as a data source should consider the technical and methodological challenges. Unfortunately, the authors do not recommend that other researchers include the discussed limitations in their research. When examining the research with Twitter data discussed in the book, it is clear that addressing limitations is not common. Twelve chapters discuss research that uses hashtags or keywords in one way or another; two examined accounts or retweets while the other chapters discuss different aspects of Twitter. Of those twelve papers, only two mentioned how the hashtags or keywords were selected; the other ten do not mention the selection criteria of the used hashtags and two papers even did not specify the used hashtags. Of the fourteen chapters, only one mentions the specific API used for data collection, while others included which tools were used. Even though it is not necessary to include this information in all types of research, it is notable how little attention is focused on the limitations of Twitter research. Based on the reported methodologies, none of the research projects can be replicated with a high level of reliability. In addition, it is important to note that none of the papers mentions the selection bias of Twitter users. The editors highlight this selection bias in the epilog:

> By their nature, lenses amplify, skew, and distort what they depict, and we must not make the mistake of taking such observations simply at face value; Twitter is no more perfect a representation of contemporary societal structures and trends than newspapers, television, or any other popular medium is able to be (Weller et al., 2014, p. 427).

As demonstrated in the previous paragraphs, some researchers do address limitations with their research, while others pay no attention to them. These limitations have been researched extensively. The following chapter discusses some of the more prominent limitations.

5. Limitations in Twitter Research

As mentioned, Twitter research continues to develop, just as the used methods do. Similar to all types of research and methodologies, Twitter research has its limitations. Several researchers describe and analyze these limitations, and a short overview will follow.

5.1 Sampling of Twitter Data/Selection Bias

The case study demonstrates that a critical selection of users is important. The decisions made have been argued for, but highlight the main limitation of big data research – the selection of data. Ekbia et al. (2015) describe how this phenomenon occurs in big data research. The selection of data, or rather the consequences of selecting certain data, is described as a selection bias. Selection bias refers to the influence of the selection of the available data (Heckman, 1979). With an increase of the use of (large) data sets in the humanities (see boyd & Crawford, 2012; Burdick, Drucker, Lunenfeld, Presner, & Schnapp, 2012; Manovich, 2011), this bias, among others factors, should be addressed and discussed as being a valuable addition to the field.

In his 1979 paper, Heckman describes selection bias and what he believes are the two causes. The first refers to the "[the] self-selection by the individuals or data units being investigated." Individuals may choose to answer a question not, or partially to alter the results. The second cause Heckman describes is "sample selection decisions by analysts or data processors [that] operate in much the same fashion as self selection [sic]" (Heckman, 1979, p. 153). In this case, researchers decide whether to include or exclude certain data gathering methods, or even parts of the data itself. Heckman's rather broad definition of selection bias can refer to many stages of research. This paper focuses only on the social media platform Twitter, on research found using certain search methods, and only examines the used methodology. Even though the decisions can be extensively discussed and debated, these decisions do influence how the research is positioned and possible limitations of the research.

With these issues in mind, researchers can work around the selection bias to reduce its influence. Researchers can specifically select data from datasets, with sound motivation as to why the selection occurred, to be able to make conclusions broadly applicable. Secondly, researchers could include motivation as to why participants or specific data was chosen, as well as who and what was omitted. This brings us to the most important requirement: Researchers have to include the limitations of their research and dataset. As Ruths and Pfeffer (2014) argue, this is not a common practice and results in a misinterpretation and misrepresentation of the real world. The researchers highlight eight points, three for data collection, and five for the used methods, which they believe researchers using social media data should address. By addressing these points, researchers will limit the influence of biases and flaws, while at the same time Ruths and Pfeffer aim to increase awareness within the field (2014).

In addition to these calls, Lomborg and Bechmann (2014) make a distinction between quantitative and qualitative research. They argue that issues with sampling are less troublesome for qualitative research compared to quantitative research. However, they do encourage researchers to

> always include a critical assessment of the sample and the in-built limitations to generalization when reporting findings. The explicit address of basic sampling biases creates transparency and thereby enhances the credibility of the empirical study (Lomborg & Bechmann, 2014, p. 260).

The influence of user self-selection, as described by Heckman (1979), has been researched with regards to social media usage by Hargittai (2015). As a communications scholar, Hargittai (2015) investigated the response of university students (n=547) of a survey regarding internet and social media use. Analysis of the answers indicates that respondents have a specific preference for social media sites, which is based on their demographic background. Hargittai concludes that in some cases, results of big data research could be the exact opposite if the sample were more representative of the population (2015, p. 74). She recommends that researchers relying on data should report the details on their sample and the limitations that result from the sample. The investigated case study aimed to contain all Twitter users tweeting in Dutch. As this is not a representative sample of the complete Dutch-speaking population, conclusions can only be made about those using Twitter.

It is also important to note that even researchers with access to all available data and sufficient financial backing can reach incorrect conclusions. When Google researchers used their data to predict influenza trends, they had company backing and access to all available data, but their model soon overestimated the actual numbers of those with influenza (Lazer & Kennedy, 2015). Lazer, Kennedy, King, and Vespignani (2014) used this case to describe possible "traps in big data analysis." The first problem they describe is the "big data hubris", the assumption that access to big data replaces traditional research and analysis. They note that it is important to address the measurements, while the validity and reliability of the data should also be addressed. An example of this trap is the study by Bruns et al. (2014) on the Australian Twittersphere. At the time of writing their article, Bruns et al. (2014) were still downloading new data, with the claim that they had enough data to outline the network (p. 5). Even so, they do not mention the impact of this decision on data measurements, validity, and reliability. The second concerns algorithms. In their analysis, Lazer et al. (2014) state that the algorithm Google used to predict influenza trends was modified by Google to accommodate its business model. This practice of modifying the platform (e.g. Twitter and Facebook) is common and as a result "[and] whether studies conducted even a year ago on data collected from these platforms can be replicated in later or earlier periods is an open question" (p. 1204). It is important to note that this limits the possibilities for researchers to repeat conducted studies on these platforms. This limit was also revealed in the case study, as some of the tweets could not be downloaded again, a mere six months after the initial study. The article by Lazer et al. (2014) concludes by providing lessons to other researchers. In particular, they address the transparency and replicability of big data research. As they remark, based on the information Google released, it would be impossible to replicate their study, even if one had

access to all available data. Going beyond the fact that it is not necessary for Google to release this information, as they are a commercial company, the replicability remains important in science. Lazer et al. (2014) also remark that researchers should understand how platforms are gradually changing. They recommend replication over a period to monitor and report on platforms, and possibly algorithms.

With regard to the case study, selection bias has had its influence. While the study aims at a complete as possible dataset, the presented data has been sampled by Twitter. This is accomplished predominantly by accessing Twitter data, as discussed in chapter 5.3. It is important to note that research on the Dutch Twittersphere is sampled within the population to those users using Twitter and tweeting in Dutch. Dutch users tweeting in languages other than Dutch, or who did not use one of the keywords, were not included in the research. In an attempt to tackle the remarks made by Lazer et al. (2014) with regards to replicability, all of the used methods are included in the appendices of this paper.

5.2 Black Box

In addition to the data sampling which influences data collection, Rieder and Röhle (2012) highlight five challenges found in Digital Methods. One of these, black boxing, refers the practice of utilizing tools and methods without being able to observe how they work, check the results, and verify the algorithms. Using tools that are a black box by nature can be present on a level of data collection, analysis, or visualization. Even when the code of those tools and methods are released, Rieder and Röhle (2012) point out multiple problems, one of which concerns the two-cultures problem, or 'code literacy'. The authors refer to the divide between researchers who are able to analyze and read the code and those who are not. Their second problem refers to the developing field of data analysis. In the development of the field, tools may be used which are in their experimental state, "in the sense that the results they produce cannot be easily mapped back to the algorithms and the data they process" (2012, p. 76). As a result, interpretations of the data by using these tools may not be possible when a more recent version of that particular tool is used. Moreover, the authors remark that it may not be possible to understand these tools on a statistical concept level. In their article, Rieder and Röhle (2012) assume researchers download data themselves or use data delivered from social media analytic companies. These companies do deliver the data ready for researchers to analyze, but as Vis (2013) notes, these companies have an intrinsic black-box nature. This is also present in the tools used in the case study. As a third party web service was relied upon for downloading all tweets, a black box was encountered. The possible algorithms and filters applied to the data were not present. Also, the editors of *Twitter and Society* remark in their introduction that the current practice of reliably measuring users or quantifying social media use is not unimportant (Weller et al., 2014, p. xxxii). They also note that current methodologies for achieving these goals are not standardized or verifiable for other researchers. This results in using tools and methods as black boxes which researchers have to trust the developers of these tools to report the outcomes correctly. Even when designing their analysis of social media platforms, researchers encounter black boxing. Researchers often have to abide by conventions, handle technological challenges, and clean their downloaded social media

data. Rieder, Abdulla, Poell, Woltering, and Zack (2015) note that often these tasks are handled by specialized tools, thus adding another layer to black boxing. The authors also note that knowledge of the technological techniques could add to producing knowledge.

The most apparent proof of using a black box in the case study is the dependence on the third party. The third party used the word list to download Twitter data. However, the exact methods are unknown. It might be possible that some data downloaded by the partner was filtered out or altered[17]. If a self-developed tool was used, one could discuss the limitations of such tool in detail and provide arguments.

5.3 API

The last major limitation discussed in the literature and present in the case study is the Application Programming Interface. A services' Application Programming Interface (API) allows interested parties to access data that a company has made public. Technological expertise makes it possible to automatically download data, sometimes even specifying the types of data. The API is what Heckman (1979) would describe as a data processor, decisions made by an automated tool. Twitter uses three different APIs, each of which has different limitations. Probably the most commonly used is the Streaming or REST API. Twitter limits access to any users' tweets to the 3,200 most recent tweets. This limitation is especially difficult when a researcher tracks a list of users over a longer period. If a user has sent more than 3,200 tweets within the period of interest, not all tweets would be downloaded. This could potentially influence the results, especially if the researcher uses a tool that does not warn that the limit has been reached. This could result in a dataset with all tweets from one user in the timeframe, to only the most recent, and possibly not spanning the complete timeframe, from another user. This was also observed in the case study, as the dataset demonstrated that some users did tweet more than 3,200 times before the end of the period[18]. A manual correction was used to give an estimation of the number of tweets these users had sent in the selected period.

Twitter has made two APIs public and free, the Streaming and Search API, while the third one is only accessible to users who pay for access. All three APIs have been subject to research, especially regarding the influence they have on the dataset. The two free APIs are limited in the amount of data they release. The Search API resembles searching Twitter via the website directly, is limited to the past seven days and focuses on relevant topics contrary to completeness[19,20]. Morstatter, Pfeffer, Liu, and Carley, (2013) compared the Streaming API tot the Firehose API (discussed below). By analyzing the same events, they concluded that the size of the dataset gathered via the Streaming API was influenced by the total number of tweets on Twitter. By their

[17] However, this is unlikely, as the relationship with the third party is good and personal.
[18] For example, one's aim is to download all tweets from a list of users in the previous month. If user X has tweeted more than 3,200 times between now and the end of last month, his tweets would not be included in the dataset.
[19] Twitter Developer Information: https://dev.twitter.com/rest/public/search (last accessed on June 17, 2016).
[20] Descriptions are correct at the time of writing. Future changes by Twitter are possible.

calculations, Twitter provides access to approximately 1% of all data at that moment sent on Twitter, regardless of the topic. Earlier, González-Bailón, Wang, Rivero, Borge-Holthoefer, and Moreno (2012) compared data downloaded via the Search API to the Stream API. They used civil demonstrations in Spain as a focus point, but do not explain why. Their data revealed that 2.5% of the tweets, 1% of unique users, and 1.3% of the used hashtags downloaded via the Search API were not present in the Streaming dataset. However, it is difficult to compare these results directly, as they only used four hashtags while using the Search API, compared to 70 hashtags in the Streaming dataset. This is due to technological limitations imposed by Twitter. The dataset from the Streaming API was almost four times larger than the Search dataset. After analyzing the dataset, González-Bailón et al. (2012) concluded that there is a strong indication that Twitter returns tweets from those users who are more centrally located within the Twitter network of the Search API, compared to Streaming API.

The third API, the Firehose API, is only accessible when a user has paid Twitter or one of their commercial partners a fee. This API provides the user access to all available tweets surrounding a subject, instead of a sample. Morstatter et al. (2013) compared the data accessible via the Firehose API to the data downloaded via the Streaming API. When comparing the numbers, the dataset via the Firehose was more than double in size based on the same parameters. Morstatter et al. (2013) could observe how Twitter handles returning 1% of the Streaming API data. When the number of total tweets increased, the number of tweets returned (around the chosen topic) by the Streaming API was reduced. Upon analysis of the two datasets, the authors concluded that the Streaming API provides a fair estimate of the top hashtags used on Twitter, but this accuracy decreases when the total number of tweets is low. In addition, the total amount of traffic influences the number of tweets returned by the Streaming API. Unfortunately, as researchers do not have access to the total number of tweets, it is difficult to make conclusions and describe the limitations of the used API. Morstatter et al. (2013) also concluded that Twitter filters the data presented in the Streaming API.

The last method for gathering data from Twitter is through commercial partners. Driscoll and Walker (2014) compared two datasets, one collected through a commercial partner from Twitter, the second sets via the Streaming API. During a short time-based event[21], a presidential debate, approximately 20% of the tweets in the commercially available data were not found in the Streaming dataset. Over a longer period, 15 days, they found that 5.2% of the tweets were missing in the Streaming dataset.

Based on these research papers, one can gain relatively good insight in how Twitter handles data sharing through APIs. The Streaming API returns a large percentage of the tweets when used in research projects that track tweets spanning multiple days, or even weeks (Driscoll & Walker, 2014, p. 1756). The Streaming API also provides a good representation of a particular event compared to the overall activity. In addition, this API also provides an accurate estimation of the

[21] Twitter aims to cover events, and gives advertisers information on the possibilities surrounding these events; blog.twitter.com/2015/introducing-event-targeting (last accessed on June 17, 2016)

top hashtags throughout Twitter, especially when the gathered dataset is large (Morstatter et al., 2013). Lastly, when comparing the Streaming API to the Firehose API, Morstatter et al. (2013) could identify 50 to 60% of the top 100 key users in the Firehose dataset. This could be an indication that the algorithms Twitter uses in their API have a preference to include the top users in the Streaming API. In conclusion, based on this research, one can have a fair indication of the results returned by the APIs, but one must remain aware of the limitations of using the Twitter API.

In addition to the limitations of Twitter APIs, researchers must also address the challenge of deleted tweets. Should they remain part of their analysis, even if the users deliberately wanted to remove their message? As Croeser and Highfield (2015) state, users even post screenshots from deleted tweets to show them to their followers. As the deleted tweets and the screenshots are included in the dataset, researchers have to make decisions about handling these tweets.

6. Completeness

In addition to the limitations discussed in multiple papers, this paper introduces another, completeness. Completeness has not been discussed before, but it does influence Twitter research. Whether a dataset is complete depends on the research question. In the discussed case study, an attempt was made to analyze all Dutch tweets in one week. As a result, a complete dataset does contain all tweets. If, for example, one is interested in analyzing a fan base, a dataset could be complete when the majority has been tracked. However, it is important to put the completeness of a dataset in perspective to the research question. The concept of completeness is discussed on three levels within Twitter research. The first is completeness in the literal sense, where a researcher has collected all available data in a dataset, and whether a researcher can claim this. The second level refers to the research methodology. It discusses the possibility of sharing the research methodology. The third level encompasses the discussion of how limitations in research are examined, and how this influences completeness. Lastly, a proposal is made on how researchers can classify their level of completeness in their Twitter study.

6.1 Conceptualizing Completeness

6.1.1 Literal Definition

The literal definition of completeness[22] in datasets refers to the notion of having collected all available data regarding the subject. In the earlier examples, it is important that a researcher has collected all available data, or at least mentions the possible limitations. However, it is almost impossible to make a claim of completeness regarding Twitter data. As one does not know the size of the total Twitter user base, and thus all messages sent, one cannot make claims about the completeness. It is, however, possible to estimate the level of completeness, as can be seen in chapter 3.1.1.

boyd and Crawford (2012) state that "bigger data are not always better data" and the size of a dataset does not say anything about how "good" the data is. They mention that understanding the sample of the data has become more important. They continue by using Twitter as an example: “Twitter does not represent ‘all people’, and it is an error to assume ‘people’ and ‘Twitter users’ are synonymous; they are a very particular sub-set” (boyd & Crawford, 2012, p. 669).

Indeed, the size of a Twitter dataset does not indicate the quality. In addition, when analyzing Twitter, researchers have to bear in mind they are (generally speaking) only analyzing messages from a small section of Twitter users, who themselves are a small portion of the general public. However, if researchers aim to analyze and make claims about all Twitter users from a specific country (Bruns et al., 2014), the most active users tweeting about a certain topic (Ausserhofer & Maireder, 2013), or for example a specific event (Ross, Terras, Warwick, & Welsh, 2011), collecting all available tweets is essential for drawing conclusions. In the example of the most active users, this definition is subject to change. Researchers have to argue that their user selection is

[22] The Oxford Dictionary defines ‘complete’ as “Having all the necessary or appropriate parts” (Oxford Dictionaries, n.d.)

representative of the population. However, they do have to download all tweets from these users and thus require a complete dataset.

6.1.2 Disclosure of Methodology

The second notion of completeness discussed is the disclosure of methodology within research. As Ekbia et al. (2015) describe, methodological issues mainly arise when researchers have to make subjective decisions. They visualize this by highlighting data cleaning and statistical significance. While researching datasets, researchers often have to clean their data; this entails deciding which elements to use for their analysis. This can range from omitting pictures, internet links, or locations to selecting only these objects. Statistical significance refers to the selection of data in a researcher's analysis. As boyd and Crawford (2012) point out, one can see "patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions" (p. 668). If a researcher chooses not to include a detailed methodology, either in the paper itself or available on request, these decisions remain hidden. Completeness of methodology also refers to explaining how the data was gathered. This includes specifying the timeframe, the used API, tools, hashtags or users, and data handling. By specifying the timeframe, one can determine whether world events, such as the Olympic Games or a terrorist attack, could have influenced the discourse on Twitter. The used Twitter API, as discussed, does influence how much and which data researchers can access. According to Bruns and Burgess (2016), Twitter specifies how many messages have been omitted when using a publically available API. If researchers include this type of information, others could judge the reliability and completeness of their research. Specifying the used tools in the research allows others to evaluate the reliability of the tools, and analyze the level of black boxing. As seen in the different research papers, not all methods of directly gathering data remain available (e.g. Kwak et al., 2010), while third party tools do not always remain available (Bruns & Burgess, 2016).

The selected hashtags or users for downloading data contain valuable information. As discussed, researchers select hashtags without specifying which words have been omitted and the reasons for their omission. In addition, tweets about the same topic but using different hashtags are omitted. This is the same for comments on tweets that do not mention the original tweet or author (Croeser & Highfield, 2015, p. 185).

In addition, some researchers state that with the use and rise of big data, theories and models are not needed, as big data enables an "empiristic mode of knowledge production" (Kitchin, 2014, p. 3)[23]. However, as Kitchin (2014) remarks, the gathering and analysis of big data is subject to interpretation and tools. If these are acknowledged, embedding the results and methods in wider debates will have more grounding. As such, this paper argues that including the methodology improves the completeness of the research, and in a sense the completeness of the dataset.

[23] Also see Kitchin (2014) for an overview of researchers stating this.

6.1.3 Addressing Limitations

The last notion of completeness within Twitter research is how the limitations are addressed. Even though it may sound logical, not all research papers include this section. As discussed, Twitter research is accompanied by several limitations. The main limitation is Twitter's API. The different APIs allow researchers different levels of access, each with their unique limitations. Because one requires significant technological skills to access this data, many researchers use specially developed tools, thus making use of a black box. If such a tool has released its source code, one has to be able to "read" the code. Another limitation that must be addressed is the selection bias. As seen, researchers using Twitter as a basis have a small selection of Twitter users, and only a small selection of the public. Completeness within Twitter research also refers to addressing the limitations of the research. As discussed further along, not all papers address these limitations, or they address them only partially.

As seen in the previous chapter, completeness is a concept influenced by the research context. More importantly, whether a dataset is complete is subjective. If a researcher has all relevant data, which is thus representative, their dataset is complete. However, others may argue that the researcher has missed relevant data. If the decisions made for collecting data are argued for, the author of the research can motivate the completeness of the dataset.

6.2 Replicability

Another limitation that researchers face is the possibility for other researchers to replicate the results. This aspect has been discussed in only a few papers. As Lazer et al. (2014) note, replicability of a study is "*a growing concern across the academy" (p. 1203).* As mentioned, even if one had all data the researcher had, it would be impossible to replicate the results based on the released methodology. Ekbia et al. (2015) makes similar assertions regarding replicability in general.

In 2013, the Australian researcher Axel Bruns described the key challenges in social media research. One of these, the replicability of the results, describes how researchers could discuss this topic (Bruns, 2013). Bruns remarks how replicability is the basis for scientific research, and thus also in social media research. He describes replicability as the use of the same methodology with a focus on a different aspect of the same phenomenon. For example, when observing an event on Twitter using one hashtag, researchers are unable to make statements beyond that event, unless others research different aspects (hashtags) using the same methodology (Bruns, 2013). In addition to describing how methodologies ensure replicability, he also notes how analysis of the same dataset by different research teams contributes to this concept. However, due to how researchers use different tools, either self-developed or readily available, in combination with the various methods to access APIs, *"there is no guarantee that two teams of researchers attempting to gather the same data at the same time will end up with identical datasets."* Bruns proposes possible solutions to these challenges. One of these is sharing methods and data between researchers. He also notes that Twitter does not allow for such practices. This could potentially result in a black market for data, as Puschmann and Burgess (2014) describe.

The ongoing development of Twitter as a platform also influences the possibilities for replicability. As Croeser and Highfield (2015) describe, with the addition of support for non-Latin scripts, such as Arabic, influences the used methodologies and the replicability. They continue by describing the difficulty of replicability in online research. With new tools that make it easier to download data without any technological knowledge, gathering becomes easier. However, as Twitter does not allow for sharing the dataset, if a researcher is not able to replicate the dataset from other researchers, it is almost impossible to replicate the results and evaluate the accuracy and value of that study.

Unfortunately, none of the discussed research in chapter 4 provides an extensive methodology or options to validate their methods. In addition, the API is not discussed in detail in these articles, even though it does significantly influence data gathering. As a result, it is difficult or even impossible to repeat the research by others, or to analyze the used methods.

7. Proposals

Based on how the limitations of Twitter research are currently discussed (chapter 4), and how completeness is influenced by the methodology, a model is proposed. This model discusses how limitations of the method used in Twitter research influence the completeness. The first level is applicable to research aimed at mapping a community on Twitter, as in the case study, the literal sense of completeness. The second level proposed is the methodology of Twitter research. The last level is the discussion of the limitations found in Twitter research. As discussed in chapter 4, these different levels are rarely discussed. These levels are discussed in reverse order, as the last level influences the other levels the most, while the second has less influence over the other two, and the last the least influence.

7.1 Addressing Limitations

The proposed frame has three layers, in accordance with the three levels of the conceptualization. The first layer describes the inclusion of limitations in research. This layer should be included in all research to enable others to put the results in perspective.

The paper on which the case study is built attempted to analyze all Dutch tweets in one week. In their report, Krijger et al. (2016) reported some limitations on their research. The case study's limitations are also addressed in this paper. However, as demonstrated in chapter 4, addressing the limitations of a study is not common practice. One of the first aspects researchers should address is how they accessed Twitter, which API they used. As the used API influences the data and thus the results, researchers should aim at choosing the most suitable API and motivate this. Both Gaffney and Puschmann (2014) and Lomborg and Bechmann (2014) provide overviews of the APIs Twitter made public as well as tools for downloading this data. Researchers should also note that it is not possible to download all tweets automatically[24] and how they handled this issue.

Within the limitations aspect, the selection bias is important. Even though most researchers acknowledge that their research is limited to Twitter itself, only some do recognize that they can track only a subsection of Twitter. Especially with hashtag-based research, the limitations of selection bias should be clear. Twitter users not using a specific hashtag, or responding to others but excluding the hashtag in their messages, will be overlooked. As a result, researchers miss larger parts of the conversation and thus context (Ruths & Pfeffer, 2014). This was proven by the need to improve the used word list in the case study, as 40.66% of the tweets did not contain a keyword and were not included in the dataset.

Thirdly, when using tools, researchers also introduce a black box. By using a predeveloped tool, researchers are unable to know what the tool does. If a researcher develops their own tools, others cannot check on the used procedures and algorithms. As a result, using tools is both a black box for the researchers as well as for the readers. It is likely that it influences the results and conclusion of the studies that researchers perform.

[24] Twitter limits the access to the last 3,200 tweets

Using tools besides a black box also introduces a two-cultures problem. This problem takes place either on a technological level, being able to understand and read a program's code (Rieder & Röhle, 2012) or on a financial level (boyd & Crawford, 2012). The financial level refers to researchers being able to buy their data and thus having access to data older than the last 3,200 tweets. As some researchers do not have the skill to read, or the financial option to buy Twitter data, it does introduce limitations to research papers. By including the specific tools used in the case study, others can assess and evaluate the options of the used methodology. The used tool in the case study had to be purchased to unlock the possibility to download the data. Even though it was not expensive[14], it does add a financial barrier, which can influence accessibility for others.

7.2 Including Methodology

The research methodology is crucial. When choosing a research design, researchers can influence the outcomes. If the methodology is not included in the paper, or only sparsely, it is not possible for others to repeat and thus determine the validity.

In the discussed papers, only two included the selection criteria for their hashtags or keywords, while two others did not even include the used hashtags or keywords. As the choice for including or excluding certain words influences the results, these should be provided. This list of the used hashtags and/or keywords should be included either in the paper itself or as an appendix. Moreover, the motivation for selecting those words is important to determine the validity. For example, when researching epidemics, it is important to know which words were not included because they do not define an epidemic. Collier et al. (2010) focused on influenza outbreaks but did not indicate which words were omitted and why the selected seven were used. If these criteria were included, other researchers would be able to assess the scope of the research and validate the results.

Secondly, it is important to describe the used methods when accessing or using tools. This includes how the data was cleaned and the processes used to analyze the data. As both of these processes alter the data, outlining the process adds another possibility for others to determine the validity of the report.

Lastly, the timeframe and the metrics of the data are relevant. These seemingly small parts of the methodology can greatly influence the Twitter dataset. As world events can influence the use and activity of Twitter[25], it can add noise to a dataset or alter the returned data of an API. Including the timeframe is an initial step to determine these points. The metrics of the data should at least include the number of tweets gathered. In addition, researchers could include the number of unique users, retweets, pictures, and URLs. This information indicates the completeness of the dataset. See Bruns and Stieglitz (2013, 2014) for an overview of the available metrics.

[25] For example, the World Cup Football in 2014; https://blog.twitter.com/2014/insights-into-the-worldcup-conversation-on-twitter.

7.3 Completeness

The last level proposed ties in closely with the previous two. The literal sense of completeness is important to address in Twitter research. Even though all research has to make a sample of the population, Twitter research has the connotation of having a complete sample. Some researchers contend that this is not the case, but especially when researching a Twittersphere, this is often not addressed. Including the selection criteria is the first step to assess the completeness of the data. For example, Bruns et al. (2014) attempted to map the Australian Twittersphere. However, due to their methodological choices, it is unlikely they included every Australian Twitter user. Even without all users, it is possible to reach conclusions based on the dataset; it remains important to disclose that the dataset is not a complete representation of all Australian users. Bruns et al. (2014) do not discuss the level of completeness of their dataset, which leaves gaps in their analysis. They could potentially have omitted a large group of Australian users, who do influence the majority, based on their selection criteria. If so, this influences their analysis and conclusions made on the Australian public.

This could be complemented by including a calculation of the completeness, as carried out in the case study. By selecting a random sample of Twitter users and comparing whether all tweets on a particular topic were included in the data sample, one can calculate the inclusion percentage. Even though the percentage is an estimation, it indicates the completeness and is relatively easy to accomplish.

Data sharing could also be included in this level. Some researchers have made their data available for others to download, but had to take their dataset offline due to changing terms of service from Twitter (Kwak et al., 2010). This has instigated a so-called black market, as described by Puschmann and Burgess (2014). Borgman (2009) describes some reasons why scholars do not share data, but also explains how it could be shared. If methods are developed to share the data and remain within the Terms of Service of Twitter, the repeatability of research is significantly easier to assess. These methods should also include anonymization, which currently is not standardized (Lomborg & Bechmann, 2014).

Lastly, authors should indicate the completeness of their data. This is particularly the case when researchers aim at analyzing at including users within specific selection criteria. As seen in the case study, an extensive list of keywords could only track just under 60% of all Dutch tweets. If researchers include a calculation, which is relatively easy to accomplish[26], others have an indication of the completeness. Again, by addressing and putting this percentage in perspective, one can validate the dataset.

7.4 Concluding the Proposals

The proposed three layers tie into each other, and are complementary at each level, as can be seen in Figure 1. Each layer improves the possibility for researchers to validate and interpret the

[26] By selecting a random sample of the Twitter users and comparing whether all tweets on a particular topic were included in the data sample, one can calculate the percentage of inclusion.

findings from Twitter-based research. This paper proposes that each research at least addresses its research limitations. If the limitations are discussed, it is likely that each scenario has been considered and the best course of action has been taken. In addition, by placing research in perspective, conclusions can more easily integrate in a scientific and public debate.

When one includes the research methodology, it is easier for others to replicate the study and to validate and assess the findings. As repeatability in research is important, as demonstrated by the Open Science Collaboration (2015) by repeating psychological experiments and reaching different conclusions, it should be more common to include Twitter research methodology. However, as noted by Bruns (2013), often journals have a word limit for articles and as a result, an extensive methodology might not be possible in the article itself. This solution is to make an extended methodology available as an appendix on the author's website or upon request by other researchers. The same would be possible for sharing the used hashtags or keywords and the dataset itself.

Completeness is the last level that researchers could include. By including the decision to limit certain aspects and providing an estimation of the completeness of the dataset, research validity is significantly improved. It is also possible to share the dataset with others, to enable them to make their own analyses based on the same dataset. As mentioned, at the time of writing no standard practices are established for sharing a sensitive dataset between researchers.

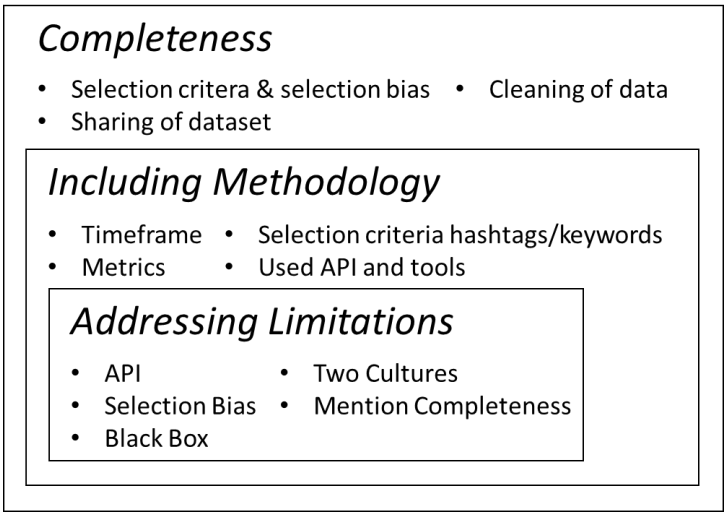

Figure 1. The three levels of completeness visualized.

8. Conclusion & Reflection

8.1 Limitations of Twitter Research

With the emergence of the Digital Methods within digital humanities, methods are being developed to analyze different digital sources. Twitter is a popular source, as it is free to access, the messages adhere to a strict model, and it is relatively easy to download data. As with all methods, there are certain limitations with the current methods for accessing Twitter as a data source. These limitations have been researched in different papers, mostly with a focus on one limitation at a time. The selection bias Twitter invokes for researchers is addressed as something to consider when using Twitter data. This ranges from how Twitter presents data to researchers to the division between researchers who are able or unable to purchase and download Twitter data. Many researchers use or develop tools to automate the process of downloading data. This introduces another limitation, as discussed by Rieder and Röhle (2012), a black box. The black box refers to the programs used, the processes of which researchers know very little. By using these programs, researchers rely on processes that they cannot monitor, and which could potentially influence the outcome. As this practice is common and not always avoidable, it is not necessary that these tools be avoided. However, it is important that this practice be addressed. These (self)developed tools make use of one of three APIs Twitter has made available. These allow downloading data directly from Twitter. However, Twitter limits the output of these APIs to prevent excessive use. As demonstrated in experiments, one of the more commonly used API, the Streaming API, is severely limited by Twitter. Researchers, however, rarely address this limitation, if the used API is mentioned at all.

The limitations mentioned earlier influence research replicability and data completeness. As discussed in this paper, these limitations are discussed only rarely or selectively by authors when describing their Twitter-based research. As these limitations influence decisions and outcomes, it is important to include a section that addresses these aspects in a research paper. This paper introduces a model based on three levels.

8.2 Making a Model

The model is based on the conceptualization of completeness in Twitter research. By introducing this concept within Twitter research, it is possible to assess the validity of a project. Ideally, all three levels would be discussed, and the choices made motivated based on these points[27]. The first level, and maybe the most important, is addressing the limitations in the research. As discussed, the limitations take place on different levels, and all influence the results. By highlighting the discussed points, researchers enable others to evaluate the research itself, as well as the results. This is particularly the case in the emerging field of the Digital Methods. As mentioned by Ausserhofer and Maireder (2013), each research develops its own methods and explores the possibilities. As a result, each project runs into different limitations that have to be addressed. The discussed points in the model, therefore, are not expansive, and others may be added or

[27] See Figure 1, p. 25, for an overview.

omitted. Unfortunately, the reviewed research papers only highlighted some or none at all, thus leaving gaps in their research.

The second level discussed research methodology. As mentioned before, each research develops a method; there is no standard practice. This results in research that is difficult or cannot be replicated. This paper argues that when the methodology is included in the paper, either in the method section or as an appendix, the validity of the research is improved. When others can replicate the research, the trustworthiness increases. Especially if researchers rely on hashtags or keywords, it is important to highlight the selection process of these words. If the argumentation for the selection is available, others can evaluate whether all angles of the phenomenon have been discussed. As discussed, in *Twitter and Society*, only two of the twelve research projects using hashtags motivate why the selected hashtags were used (Weller et al., 2014). However, as all the downloaded data, and consequently the results, do rely on these words, it is important that the motivation is clear. Ideally, researchers would outline all the used methods systematically to enable others to replicate the research.

The third and last level of the model discusses the completeness of the dataset. This is especially interesting for those projects attempting to discover and analyze Twitterspheres. These projects consist of determining a focus group, finding them, and analyzing their behavior. For example, attempts have been made to analyze entire countries (Aslanyan & Gillespie, 2012; Ausserhofer & Maireder, 2013; Bruns et al., 2014) or political parties (Paßmann et al., 2014). In these cases, it is important to know how complete the dataset is compared to all available data from Twitter. The case study in this paper examined how complete the dataset of the Dutch Twittersphere was. By analyzing a random sample of all users, it was determined that the complete dataset consisted of 59.34% of all Dutch tweets. The developed method could be used by others or used as a basis for a different method of analysis.

8.3 Reflection on Research

This paper is not the first to describe how researchers could improve their papers by including methods or limitations. Kitchin (2014) describes the different aspects researchers have to consider when using big data. Ruths and Pfeffer (2014) introduced a model that researchers could use when collecting and analyzing data, and concluded that researchers should have an increased awareness of what and how exactly data is analyzed. With their overview of Twitter research, Williams et al. (2013) discovered that over 80% of the papers did not include any quantitative information in the abstract, based on 575 papers. They recommend that authors include quantitative information in the abstract. With regards to services' API, Rieder et al. (2015) recommends acknowledging API limitations and placing the results in perspective. Interestingly, Bruns discusses the limitations of big data research in 2013[28], and recommends steps to be taken.

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

However, his papers published after this article in *First Monday*, do not address the discussed limitations, or provide a comprehensive methodology.

This paper proposes a model that others can use in their Twitter-based research when addressing the limitations of their research as well as the completeness of their dataset. It combines and complements previous research that addresses the limitations, and proposes a comprehensive model. When examining previously published papers, especially those using hashtags, several pitfalls and limitations are discovered. Unfortunately, not all papers related to Twitter-based research include the number of collected tweets, which hashtags were used, or the possible limitations of the research. These points have been addressed and discussed, and as a result, the model has been developed.

The proposed model is a start, and possibly not complete nor comprehensive. It provides an overview of the limitations that influence Twitter-based research, which should be addressed. It can also be used to develop a standardized research method research, which would contribute to digital humanities and Digital Methods.

8.4 Future Research

This paper focused on the limitations of Twitter research, in particular the methodologies and completeness of datasets. The proposed model was designed based on previous research discussing limitations, as well as research based on Twitter data, which lacks the discussion of limitations. However, it does leave room for future research.

A major aspect that needs to be researched is methods standardization within Twitter research. Even though different aspects are studied, one base guideline could improve the readability, validity, and sharing of research. This standardized guideline should also address the limitations discussed in this paper, provide alternative tools, or highlight the discussion about these limitations. This will possibly enable researchers to make conscious design decisions.

Another aspect relevant to the previous point is developing a method for sharing and anonymizing data between researchers. As discussed, it is relatively difficult for researchers to access and download data, and even more difficult to download the same data. If researchers are able to analyze the same data others had access to, it would possibly highlight different aspects of the same phenomenon.

Lastly, a topic relevant to Digital Methods has not been discussed in this paper. The ethical side of using social media data is still subject to research. As this is a different side of the discussion, the ethical concerns were not addressed in this paper. However, it should be noted that most of the discussed papers mention the ethical aspect. This research topic of research could be implemented in the previously mentioned models.

9. Appendix

9.1 Appendix I: List of Figures

9.2 Appendix II: Randomly Selecting Users

This appendix highlights the methodology for selecting random users. This is based on the file with all tweets of one week.

Firstly, all users of the datasets were isolated and filtered to remove duplicates. This list of users was analyzed using Microsoft Excel 2013. Each row contained one user, not sorted in any manner.

In a column next to a name, Excels "=RAND()" function was used to calculate a random number between 0 and 1. This number was made static to prevent changing in later steps. As a result, each name had a linked random number, as seen in Table 1.

Table 1

Example for name and random number combination

Username	Number
Name1	0,611573
Name2	0,040538
Name3	0,111295
Name4	0,741337
Name5	0,822197
Name6	0,762631
Name7	0,08559

The column containing the numbers was sorted, either from small to large or vice versa. The column with the names are automatically linked together, and thus will be sorted in a random order, see Table 2.

Table 2

Example for name and random number combination sorted by number

Username	Number
Name5	0,231229
Name6	0,34462
Name2	0,546681
Name7	0,753871
Name4	0,848677
Name3	0,852966
Name1	0,893911

Of the randomly sorted list, the first 50 users were selected which resulted in 400 usernames. These usernames were selected to download all tweets from.

9.3 Appendix III: Downloading All Tweets

After the random selection of the users several methods for downloading all of their tweets were considered. On method relied on directly accessing Twitters API to download the tweets. However, this proved to be a sophisticated technological task. Third party tools were considered, but most of these were either offline, not working or needed large computational power.

A combination of two websites and two computer programs were used. The selected accounts were put in a list in Microsoft Excel 2013, each name on a new row. The webservice Twitonomy[29] offers registered Twitter users the possibility to analyze Twitter accounts on their metrics, as well as downloading all tweets. This download is presented in an Excel file, which can be used to analyze. To automate this process, a small computer program was used to automate the needed steps[30]. This program enables users to record the steps taken with keyboard and mouse, and allows them to be repeated.

To select only the tweets in the investigated week, a macro was used. Macro's are small series of instructions which Excel will automatically perform. By combining a macro which filtered out all tweets outside of the selected week, and a macro which automatically performs this task on all Excel files within a folder, the correct tweets were filtered and counted. This process was repeated for the data files consisting of all tweets of one week. By filtering based on usernames, the total number of tweets captured could be determined.

Finally, the number of tweets by the random sample of users, and the number of tweets in the original dataset were compared. After filtering out those account in a different language it was shown that 59,34% of all tweets were captured.

[29] http://www.twitonomy.com/, last accessed on 14 June 2016
[30] Tinytask, available for free at http://www.vtaskstudio.com/support.php, last accessed on 14 June 2016

The used macros are included without expansive notes.

Sub ProcessFiles3()
Application.ScreenUpdating = False
  Dim Filename, Pathname As String
  Dim wb As Workbook
 	
  Pathname = "F:\Twitter Data\Files\"
  Filename = Dir(Pathname & "**.xlsx")
  Do While Filename <> ""
  Set wb = Workbooks.Open(Pathname & Filename)
  V2DateSplitAdvancedFilterV2 wb
  wb.Close SaveChanges:=True
  Filename = Dir()
  Loop
End Sub
	"Sub ProcessFiles3()" runs through all files within a folder, and applies "V2DateSplitAdvancedFilterV2" Taken form https://stackoverflow.com/questions/14766238/run-same-excel-macro-on-multiple-excel-files

Sub V2DateSplitAdvancedFilterV2(wb As Workbook)
  With wb
' V2DateSplitAdvancedFilterV2 Macro
'
Dim GetBook As String
Dim lastRow As Double
GetBook = ActiveWorkbook.Name
'
' Date Split
  Columns("B:B").Select
  Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
  Range("A4").Select
  Range(Selection, Selection.End(xlDown)).Select
  Selection.TextToColumns Destination:=Range("A4"), DataType:=xlDelimited, _
  TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=False, _
  Semicolon:=False, Comma:=False, Space:=True, Other:=False, FieldInfo _
  :=Array(Array(1, 4), Array(2, 2)), TrailingMinusNumbers:=True
' Preparing advanced Filter
  Range("A4").Select
  Range(Selection, Selection.End(xlToRight)).Select
  Selection.Copy
  Range("L4").Select
  ActiveSheet.Paste
'Select date range
 	
Windows("CountOfValuesWithinData.xlsx").Activate
	"Sub V2DateSplitAdvancedFilterV2" is inherited from "ProcessFiles3()" and performs the macro specified below
	
	These rows select all data and transforms text to colomns. Subsequently, rows are inserted and the date & time information in the file is split. Lastly, the data is filtered so only tweets in the selected time frame are selected. The total number of tweets within the selected week are copied to a different Excel file "CountOfValuesWithinData.xlsx" where the information of each user is stored on a new line.

```
    Range("A2:A9").Select
    Application.CutCopyMode = False
    Selection.Copy
    Windows(GetBook).Activate
    Range("L5").Select
    ActiveSheet.Paste
    Range("E8").Select
    Application.CutCopyMode = False
'Advanced Filter
    Range("A4:J3193").AdvancedFilter
Action:=xlFilterCopy, CriteriaRange:=Range _
        ("L4:U12"), CopyToRange:=Range("L16"),
Unique:=False
    Range("L15").Select
    ActiveCell.FormulaR1C1 =
"=COUNTA(R[2]C:R[2985]C)"
    Application.Wait (Now + TimeValue("00:00:1"))
    Range("L15").Copy


Windows("CountOfValuesWithinData.xlsx").Activate

        Dim ws As Worksheet
    Set ws = ActiveSheet
    For Each cell In ws.Columns(1).Cells
        If IsEmpty(cell) = True Then cell.Select: Exit For
    Next cell
    Selection.Value = GetBook
    Selection.Offset(, 1).Select
    ActiveSheet.Paste
    Selection.PasteSpecial Paste:=xlPasteValues,
Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False
    ActiveWorkbook.Save
    Windows(GetBook).Activate
    Application.Wait (Now + TimeValue("00:00:01"))
    End With
End Sub
```

9.4 Appendix IV: Glossary

API / Application Programming Interface

A method which developers and researchers can use to access databases. Social media platforms make these available to interested parties to download data and perform analyzes.

Black box

A program which accepts data as input, performs calculations and provides an output. The calculations are not known and cannot be checked.

Dataset/Database

A digital file containing information. Commonly structured in a specific format. Can include a large quantity of data or only one row of data.

Followee

A user who is followed by certain users on Twitter

Follower

A user on Twitter who follows a certain user on Twitter. Receives all information send by that particular user

Hashtag

A specific keyword on Twitter which is preceded by a hashtag sign (#). Allows users to track, contribute to or search for a specific topic

Tool

A different name for a computer program

Tweet

A message send on social media website Twitter. A tweet has a maximum of 140 characters and could include mentions, pictures, URLs and hashtags.

Twitter

Social media platform which allows users to post messages consisting of 140 characters. Can include pictures, videos and internet URLs.

Twittersphere

All users of a specific country, culture or group on Twitter. Is mostly referred to when analyzing users from a specific country.

10. Bibliography

Aslanyan, A., & Gillespie, M. (2012). The Russian-language Twittersphere, the BBC World Service and the London Olympics. *Participations, Journal of Audience & Reception Studies*, *12*(1), 608–629. Retrieved from http://participations.org/Volume 12/Issue 1/34.pdf

Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Information, Communication & Society*, *16*(3), 291–314. http://doi.org/10.1080/1369118X.2012.756050

Berry, D. M. (2012). Introduction: Understanding the Digital Humanities. In *Understanding Digital Humanities* (pp. 1–20). London: Palgrave Macmillan UK. http://doi.org/10.1057/9780230371934_1

Borgman, C. L. (2009). The digital future is now: a call to action for the humanities. *Digital Humanities Quarterly (DHQ)*, *3*(4), 1–30. Retrieved from http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

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

Bruns, A., & Burgess, J. (2016). Methodological Innovation in Precarious Spaces: The Case of Twitter. In *Digital Methods for Social Science: An Interdisciplinary Guide to Research Innovation* (pp. 17–33).

Bruns, A., Burgess, J., & Highfield, T. (2014). A "Big Data" Approach to Mapping the Australian Twittersphere. In P. L. Arthur & K. Bode (Eds.), *Advancing Digital Humanities: Research, Methods, Theories* (pp. 113–129). Palgrave Macmillan. http://doi.org/10.1007/978-0-85729-493-7_78

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: metrics for tweeting activities. *International Journal of Social Research Methodology*, *16*(2), 91–108. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13645579.2012.756095

Bruns, A., & Stieglitz, S. (2014). Metrics for understanding Communication on Twitter. In *Twitter and Society* (pp. 69–82).

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012). *Digital_Humanities*. Cambridge, Massachusetts: The MIT Press. http://doi.org/10.1108/S2044-9968(2013)0000007006

Collier, N., Son, N. T., & Nguyen, N. M. (2010). OMG U got flu? Analysis of shared health messages for bio-surveillance. In *CEUR Workshop Proceedings* (Vol. 714, pp. 18–26).

Croeser, S., & Highfield, T. (2015). Mapping Movements – Social Movement Research and Big Data: Critiques and Alternatives. In G. Elmer, G. Langlois, & J. Redden (Eds.), *Compromised Data: From Social Media to Big Data* (p. 296). Bloomsbury Publishing USA.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and

Twitter. *PLoS ONE*, *6*(12).

Driscoll, K., & Walker, S. (2014). Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, *8*(1243170), 20. Retrieved from http://ijoc.org/index.php/ijoc/article/view/2171

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., … Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*. John Wiley and Sons Inc.

Gaffney, D., & Puschmann, C. (2014). Data Collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 55–68). Peter Lang.

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *"Raw Data" Is an Oxymoron* (pp. 1–14). MIT Press. http://doi.org/10.1080/1369118X.2014.920042

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012). Assessing the bias in communication networks sampled from twitter. *arXiv Preprint arXiv:1212.1684*, *44*(0). Retrieved from http://arxiv.org/abs/1212.1684

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://doi.org/10.1177/0002716215570866

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161. http://doi.org/10.2307/1912352

Israel, G. D. (1992). Determining Sample Size. *University of Florida, IFAS Extension*, *PEOD6*(April 2009), 1–5. http://doi.org/10.4039/Ent85108-3

Jones, S. (1999). *Doing Internet Research: Critical Issues and Methods for Examining the Net*. *Critical Studies in Media Communication* (Vol. 18). http://doi.org/10.1080/07393180128095

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1). http://doi.org/10.1177/2053951714528481

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12. http://doi.org/10.1177/2053951714528481

Krijger, R., Leeftink, J., & Brits, L. (2016). Lokaal Nederland op Twitter.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a Social Network or a News Media? *The International World Wide Web Conference Committee (IW3C2)*, 1–10. http://doi.org/10.1145/1772690.1772751

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*, *949*(February 2001), 4.

Lazer, D., & Kennedy, R. (2015). What We Can Learn From the Epic Failure of Google Flu Trends. Retrieved May 19, 2016, from http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big

Data Analysis. *Science*, *343*(6167), 1203–1205. Retrieved from http://www.sciencemag.org/content/343/6176/1203

Lomborg, S., & Bechmann, A. (2014). Using APIs for Data Collection on Social Media. *The Information Society*, *30*(4), 256–265. http://doi.org/10.1080/01972243.2014.915276

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities*, 1–10. Retrieved from http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Miller, C., Ginnis, S., Stobart, R., Krasodomski-Jones, A., & Clemence, M. (2015). *The road to representivity, A Demos and Ipsos MORI report on sociological research using Twitter*. London.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM*, 400–408. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPDFInterstitial/6071/6379

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. http://doi.org/10.1126/science.aac4716

Oxford Dictionaries. (n.d.). complete.

Paßmann, J., Boeschoten, T., & Schäfer, M. T. (2014). The Gift of the Gab: Retweet cartels and gift economies on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. 331–344). New York: Peter Lang.

Puschmann, C., & Burgess, J. (2014). The politics of Twitter data. *Twitter and Society*, 43–54. http://doi.org/10.2139/ssrn.2206225

Rieder, B., Abdulla, R., Poell, T., Woltering, R., & Zack, L. (2015). Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said." *Big Data & Society*, *2*(2), 2053951715614980. http://doi.org/10.1177/2053951715614980

Rieder, B., & Röhle, T. (2012). Digital methods: Five challenges. In *Understanding Digital Humanities* (pp. 67–84). Palgrave Macmillan UK.

Rogers, R. (2009). The End of the Virtual: Digital Methods. *Media*, 1–25. http://doi.org/10.5117/9789056295936

Rogers, R. (2013). *Digital Methods*. Cambridge, Massachusetts: The MIT Press.

Rogers, R. (2014). Debanalising Twitter: The Transformation of an Object of Study. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and Society* (pp. ix – xxvi). New York: Peter Lang. Retrieved from http://hdl.handle.net/11245/1.416833

Rosenberg, D. (2013). Data before the fact. In *Raw data is an oxymoron* (pp. 15–40). The MIT Press.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: conference Twitter use by digital humanists. *Journal of Documentation*, *67*(2), 214–237. http://doi.org/10.1108/00220411111109449

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. http://doi.org/10.1126/science.346.6213.1063

Twitter. (2016). Twitter Milestones. Retrieved March 27, 2016, from https://about.twitter.com/company/press/milestones

Vis, F. (2013). A critical reflection on big data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10).

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). Twitter and Society. In *Twitter and Society* (p. 447). Retrieved from http://www.peterlang.com/index.cfm?event=cmp.ccc.seitenstruktur.detailseiten&seitentyp=produkt&pk=71177&cid=5&concordeid=312169

Wibberley, S., Reffin, J., & Weir, D. (2014). Method51 for Mining Insight from Social Media Datasets. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations* (pp. 115–119).

Williams, S. A., Terras, M., & Warwick, C. (2013). What people study when they study Twitter. *Journal of Documentation*, *69*(3), 1–74.

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. http://doi.org/10.1108/AJIM-09-2013-0083