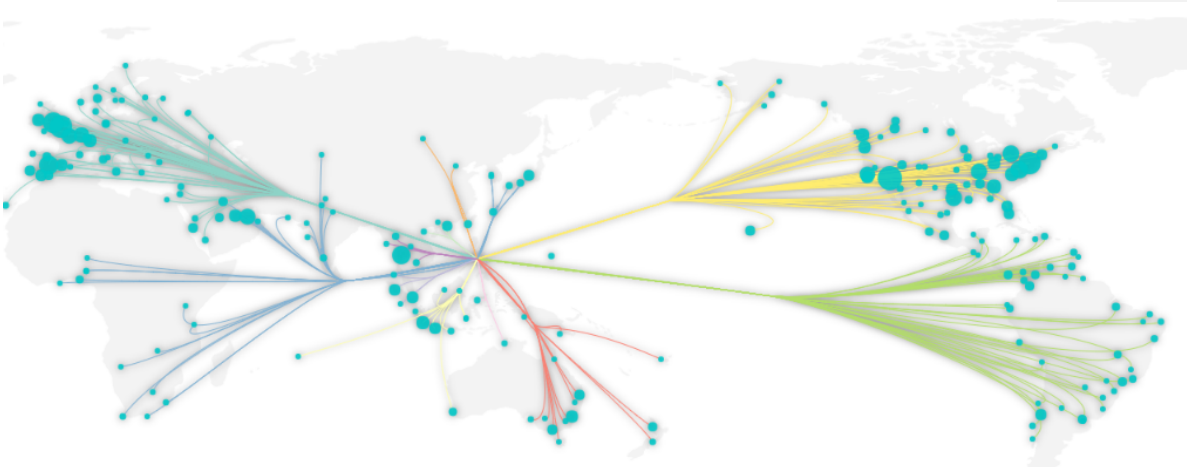
Turning Snippets into Stories:

The Potential and Challenges of Twitter Data for Humanities



Joery Kragten

S1924095

MA Book & Digital Media Studies

Supervisor: Dr. P.A.F. Verhaar

Second Reader: Dr. P.A. Vierthaler

Date of Completion: 19 November 2017

Word count: 19.800

Table of Contents

Abstract

With 300 million monthly active users, Twitter is a global social media platform embedded in everyday communication and information diffusion. As a result, it has attracted a wide range of scholarly disciplines, studying its data, resulting in hundreds of studies that utilized Twitter's data. This thesis will focus on the *challenges* and *potential* of using Twitter data for Humanities. As every social media platform has its unique dynamics, Twitters structure will be explored to see how it relates to the research questions applied to it. The process of increasing policy-orientated measures, will be described to illustrate how Twitter data is a valuable, and therefore limited accessible form of information. A number of studies is analyzed to explore potential methods and the results. Finally, a case study will be executed to demonstrate both the challenges and potential for humanities students. The focus will be on how Twitter's 140 character long tweets can eventually be used to attribute to the greater stories written about human society and culture.

Introduction

> 'Bird chirps sound meaningless to us, but meaning is applied by other birds. The same is true of Twitter: a lot of messages can be seen as completely useless and meaningless, but it's entirely dependent on the recipient.'[1]

With this quote Jack Dorsey, CEO and co-founder of Twitter, explained his choice for the bird as a logo for the social media platform.[2] Twitter was founded in 2006. In its beginning years it had the 'banal' reputation of a social media platform where friends could share their whereabouts and activities. Initially, this resulted in a platform on which the majority of the messages gave information on the composition of breakfast, lunch or dinner. Nowadays, twitter messages ('tweets') often make their appearance in the news. As the world became more and better connected, both professional and citizen journalist now utilize the platform to report on breaking news and topical events. Politicians use twitter to share their ideas and opinions, with the American president Donald Trump as one of the prime exploiters of the platform. However, the most important aspect might be that Twitter gives individuals the opportunity to 'stamp' their message with a hashtag. Twitter's hashtag can be considered as the most influential typographic innovation of the 21st century, allowing users to debate on a global level.[3]

With 140 characters[4] allowed per message, Twitter is considered to be a 'micro-blogging' platform. The length of each message was determined by the limit 160 characters of SMS for mobile phones, leaving 20 characters for the authors name per tweet. This length is particularly useful for mobile messaging. As smartphones became common good, and mobile network grew bigger and faster, the number of tweets sent per day kept growing.

---

[1] D. Sarno, 'Twitter creator Jack Dorsey illuminates the site's founding document. Part I', *Los Angeles Times*, 18 February, 2009 <http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html> (18 September, 2017).

[2] The image at the title page visualizes the worldwide response to the typhoon Haiyan in November 2013 and 'shows every geotagged Tweet mentioning the word 'help' (in 22 different languages) combined with key terms around the disaster'. Twitter Interactive, 'Philippines'<http://twitter.github.io/interactive/philippines/>(19 November, 2017).

[3] This is best demonstrated in the recent #MeToo movement, where victims of sexual violence and/or intimidation share their stories, creating global consciousness. This article explains how this movement started and grew to a global phenomenon. N. Khomani, '#MeToo: how a hashtag became a rallying cry against sexual harassment', *The Guardian,* 20 October, 2017 <https://www.theguardian.com/world/2017/oct/20/women-worldwide-use-hashtag-metoo-against-sexual-harassment>(14 November, 2017).

[4] Twitter is currently in the process of expanding to 280 characters per message for all languages except Japanese, Chinese and Korean. As this thesis was written during the time Twitter enrolled the 280 character limit, all studies mentioned here focus on the original 140 character limit. Therefore throughout this whole thesis there will be spoken of 140 character limit. A. Rosen and I. Ihara, 'Giving you more character to express youself', *Twitter Blog,* 26 September, 2017 <https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html>(14 November, 2017).

While in 2009 only 2.5 million tweets were sent per day, today 500 million tweets per day are sent.[5] The enormous bulk of tweets have attracted not only journalists and trend watchers, but also scholars, who tempt to answer a varying range of question with the help of Twitter. As Dorsey stated, the message's meaning can be determined by the recipient. Scholars can manually read Trump's tweets, or use computational methods to analyze and sort the many thousands of reactions that he evokes with his tweets. Even the banal tweets that merely expressed what someone had for lunch have proven their value for scientific research, particularly in analyses of large corpora of tweets. By comparing the use of *soda, pop* and *coke* in relation to the geo location it was tweeted from, Twitter data scientist Edwin Chen studied regional variation in language.[6]

Between the 'chirps' of one president or the 'singing' of the masses on Twitter there is a wide variety of opportunities for scientist to study information exchange, social networks, political debates and sentiments, trending topics and language use, among others. Twitter provides datasets for scholars in a wide range of disciplines, including computer and information science, communication, economics, social and behavioral sciences, and the humanities. Up to 2014 this has resulted into 380 publications in which Twitter data were utilized.[7] The datasets that are being analyzed range from a handful of tweets to some numbered in the billions. Therefore, close reading is not always an option and computational methods have to be used to count, sort and analyze the tweets. But before tweets can be researched, they have to be accessible. In order to do so, Twitter provides both researchers and practitioners a free Application Programming Interface (API) which allows them to gather and analyze large data sets of tweets. Apart from 140 characters of text, the API also provides 160 sorts of metadata extracted from the tweets or twitter users.[8] Having a greater number of metadata than characters per tweet provides researchers with a large range of opportunities, but also many limitations and challenges.

Among all disciplines in the 380 publications analyzed, disciplines related to the humanities only make up a small part of all studies. This is noteworthy because the

---

[5] Internet Live Satistics, 'Twitter Usage Statistics'<http://www.internetlivestats.com/twitter-statistics/>(17 November, 2017).

[6] E. Chen, 'Soda vs. pop with Twitter' <http://blog.echen.me/2012/07/06/soda-vs-pop-with-twitter/>(14 November, 2017).

[7] M. Zimmer and N.J. Proferes, 'A topology of Twitter research: disciplines, methods, and ethics.' *Aslib Journal of Information Management*, 66(3), 2014, pp.250–261.

[8] W. Wolny, 'Knowledge Gained from Twitter Data', *Annals of Computer Science and Information Systems,* 8, pp. 1133-1336.

humanities study ‘aspects of human society and culture’.[9] Twitter (and other social media), offers unique possibilities to study these aspects, because for the first time in history enormous amounts of personal expression are documented in the form of data. Whereas journalism has been described as ‘a first draft of history’, Twitter can be considered as a ‘first draft of the present’.[10] Whereas Twitter is gaining popularity as a platform for scholarly research[11], an overview of how Twitter can be utilized in the humanities does not yet exist. Therefore the goal of this Thesis will be to explore both the *potential* and the *challenges* for studying Twitter data in humanities.

In order to explore both the *potential* and *challenges* for humanities, firstly Twitter’s structure and history will be studied in chapter 1. As it is necessary to understand the dynamics of communication on Twitter to form potential research questions, the first paragraph will explain the main operators of Twitter. These operators formed the basis of the type of research performed by researchers, which will be the focus of the second paragraph. In the last paragraph the humanities will be explained more broadly, along with their more data-orientated counterpart: the digital humanities. Chapter 2 will focus on the technical and policy-orientated *challenges* to retrieve a sufficient data set from either Twitter’s API, commercial data sellers or public institutions like the Library of Congress. How this data is studied, which methods are applied to it and results are gained from it, will be the focus of Chapter 3. This chapter will explore the *potential* of a number of studies, to see how they either *extract* knowledge suitable for humanities, *apply* knowledge from humanities to Twitter data or combine these two approaches. In the last chapter a case study will be executed to explore the question to what extent a master student in the humanities can effectively use Twitter data. To do so the thesis will combine studies regarding the #BlackLivesMatter movement with an own study working with data gathered from #BlackLivesMatter. Finally, a general discussion will evaluate the necessity for big or small datasets. This thesis will conclude with a conclusion which sums up all *potential* applications and *challenges* of

[9] The Humanities have many definitions, this thesis will work with this short and manageable definition. OED Online, ‘humanity,n.’ <http://www.oed.com.ezproxy.leidenuniv.nl:2048/view/Entry/89280?redirectedFrom=humanities#eid311537170 >(15 November 2017).
[10] A., Bruns and K. Weller, ‘Twitter as a First Draft of the Present – and the Challenges of Preserving It for the Future’, *Preceedings of ACM Web Science Conference,* (2016) <http://dx.doi.org/10.1145/2908131.2908174 >(7 November, 2017).
[11] Scholarly works like *Twitter and Society* (2014) provide an overview of various studies working with Twitter data. K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Twitter data. Additionally, it offers a recommendation for the steps humanities scholars ought to take to adequately use Twitter data.

Thus, this thesis will provide a general overview for those active or interested in the (digital) humanities, who wish to utilize Twitter data for academical purposes. Therefore all aspects of Twitter, and Twitter data, will be explained from scratch. In the end the purpose is to explore how Twitter's 'snippets' can fit into the greater 'stories' that are traditionally written within the humanities.

Chapter 1: Twitter's History and Foundation: Structural Development, Evolution of Twitter Research and the Crossroads with Digital Humanities

Nowadays Twitter has over 300 million active users and its hashtag has become a renowned typographical element. This chapter describes the founding and history of the platform and the impact of advancing internet technology and user innovation on Twitter. First the structure will be explained, thereafter the type of research that emerged from Twitter structure and use. The third paragraph explains how Twitter relates to the (digital) humanities, what the humanities are and how it contributes to- and profits from Twitter research.

1.1 Structure of Twitter

Twitter's structure is defined by its initial purpose, sharing your whereabouts with your friends. The social relationships on Twitter are asymmetrical though, as users don't agree to a mutual friendship, but can individually choose to follow each other, resulting in unidirectional and bidirectional relationships. Followers will be provided access to a user's stream automatically, unless they have a private account. This distinguishes Twitter from other social media platforms like Facebook, where relationships are mainly reciprocal. Through the World Wide Web or with smartphones, users can dispend 140 character long tweets. In adjustment, users can add or embed multimedia content like pictures (possibly of texts), (live) videos and audio to their message. If a tweet embodies merely 'blank' text (without any @mentions or #hashtags), it will reach only the feed of a user's followers, but not necessarily those who are followed by the users because relationships are often unidirectional. The larger the number of followers a user has, the bigger the public that is reached. The personal follower network is called a 'meso' network, conforming to the layered model of communicated spaces created by Bruns and More.[12] This model displays how the use of Twitter's operators, influences the type of conversation that is held and seize of the public that is reached. The model distinguishes three levels in a pyramid form, going from the smallest public on an interpersonal 'micro' level, to a medium networked public on a meso level and a global 'macro' level where potentially any Twitter user can be part of the public which is reached.

Twitter's power is that on this macro level, it offers the opportunity for every user to reach out to and debate with a worldwide public with the help '#' symbol, called a hashtag. Added to the first letter of a word, it marks it as a keyword or topic in a tweet. A sequence of words can also be added to the hashtags, as long as no spaces are used. The tagged word or

[12] A. Bruns and H. Moe, 'Structural Layers of Communication on Twitter', in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014), pp. 15-28, p20.

sequence becomes a hyperlink, which when clicked, redirects the user to a feed that collects all tweets containing the same hashtag. The hashtag symbol was suggested by users and adopted by Twitter. The hashtag was not invented by Twitter nor its users though, as hashtags were already being used on early chats services like Internet Relay Chat (IRC), which combined them with descriptive names, or tags, to organize groups.[13] Transferred to Twitter, it became a pervasive technology, that allowed its users to publicly debate a certain topic at a 'macro' level. Recently the hashtag celebrated its tenth anniversary, of course by showing its function, offering the opportunity to filter all tweets regarding the anniversary with the hashtag #10thanniversary. Likewise, comments on sport events or on breaking news events, can effectively be clustered using the hashtag technology. It is a powerful filter that provides users with a stream of information on a specific event. Often, these public discussions rapidly form and dissolve around these kind of topical events, granting them temporary visibility in Twitter user feeds as trending topics. Generally, hashtags itself become trending topics, occasionally names, slogans and places do too, but simply because they are named in relation hashtags which are trending already.

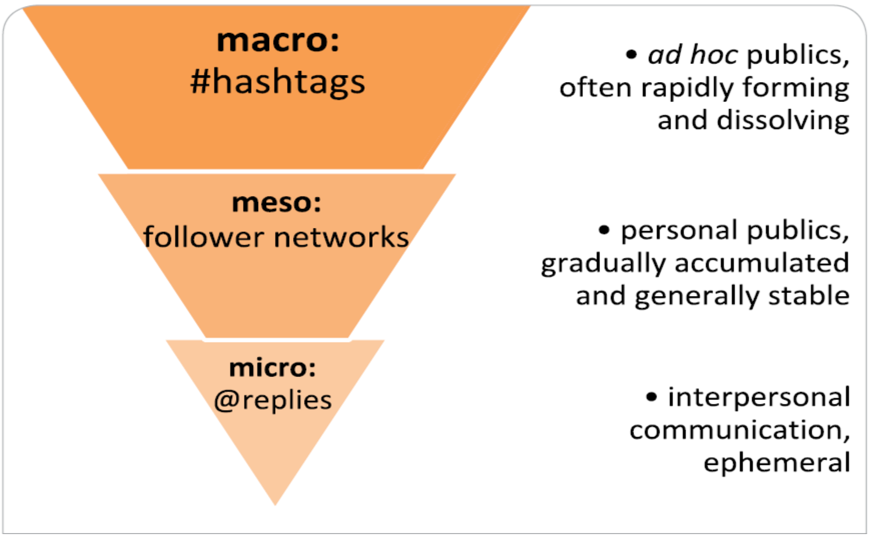


Figure 1: Layered Model of Communicative Spaces on Twitter by Bruns and More. Source: *Twitter and Society* (2014)

[13]K. Collins, 'The 10th anniversary of the hashtag is a reminder that Twitter's best features came from outside the company', *Quartz*, 24 August, 2017 < https://qz.com/1060789/the-10th-anniversary-of-the-hashtag-is-a-reminder-that-twitters-best-features-came-from-outside-the-company/>(18 September, 2017).

Apart from the hashtags, Twitter has three other operators: @ for addressing or mentioning, *http://* for linking and *RT* for republishing. The @replying operator initiates communicative interaction and awareness around users. Combined with a username, an @reply at the beginning of a tweet directly addresses this person. If the @reply is placed somewhere else in the tweet, it indirectly mentions the person. Therefore, replying and mentioning serve different strategies. A reply can be used to comment on someone to open an interpersonal conversation at a 'micro' level. Mentioning is more suitable to create attention around a user, for example to express an opinion about the referred user. In both cases, the username followed by the @ operator becomes aware of being addressed or mentioned, and has the opportunity to respond. The users following the initial tweeter become aware of persons which are mentioned, offering them the opportunity to join the conversation, follow the user or be informed about them. Followers only see the @reply tweets, however, if they also follow the person that is addressed by it. The @ operator was, like the hashtag, initiated by users. Initially, followers had the opportunity to see all @replies in their feed. When Twitter disabled this option, the users responded the day after with the hashtag #fixreplies. It became the top trending hashtag on Twitter. It is an example of how a user-generated technology like the hashtag is utilized to improve another user generated operator, the @reply.[14] The latest addition to replying is the opportunity for a user to reply to their own tweet. The @reply including their own username can be deleted to create extra space. This way, a user can make a so called 'thread', consisting out of multiple consecutive messages. Thus, the 140 character limit only fragments a thread.[15]

The third operator, the retweet (RT), owes its abbreviation to its initial use. Twitter users added RT in their tweets followed by @name and the complete copy of a user's tweet. This was the most common formulation, although 'via' or 'by' were used too. Less frequently MT was used to indicate a modified retweet. Because Twitter's experience with adapting earlier operators had learned them to take user suggestion into consideration, the retweet was carefully implemented by Twitter. The end result was a retweet button which automatically reposted a Tweet, taking RT out of the text itself. The reactions to this adaptation were mixed. Positive reactions praised the effectiveness and ease of the new button. Negative reactions criticized the inability to add comments, to set a context, to shape diffusion and to preserve

[14]A. Halavais, 'Structure of Twitter: Social and Technical', in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014), pp. 29-41, p34.
[15] Twitter, 'Twitter Support', <https://twitter.com/twittersupport/status/442433903546994688>(18 September 2017).

deleted tweets.[16] Eventually, both sides could be satisfied. Retweeting reached a larger audience and soon the option to add a comment in the form a full tweet was added. In the manual retweet the RT @name resulted in the loss of character space. Nowadays, 140 characters can be added to the retweet, leaving no reason to manually retweet. Hereby retweeting became a quick opportunity for sharing information and distributing messages to reach many people with the click of a button. The manual RT was initially used by a fairly small group of 'geeks and new folks', but paved the road for mass use in popular culture.[17]

The last operator, the hyperlink, enables users to extend the textual limit of a tweet by adding photos, videos, music or links to products, news or blogs. These additions will be displayed underneath the text message. It's especially useful to elaborate on a message, since 140 characters are too short for in depth stories. This way Twitter can help to diffuse news articles, speakers who post their videos on YouTube or photographers who post on Instagram. Posting images can also be used to work around the 140 character limit, since user can attach images of a blog. Adding multimedia content offers great opportunities to report from the ground. Users whom are present or near demonstrations, (natural) disasters, terroristic attacks or sport events, can add photos, videos or livestreams to report on the ongoing events. Again, these functions are due to the user community, in this case the more technical users. Twitter was deliberately open to alternative user interfaces via its API. Therefore, users were able to contribute to Twitter via third party applications. As a result, every popular online media platform could be quickly integrated into Twitter.[18] Vice versa, the content distributed via Twitter could easily be incorporated into other platforms.

Overall, Twitter contained a very easy to use interface, which stayed close to its original format. It kept a small number of operators, which were initiated by the user community. Twitter's accessible structure is one of the main reasons it has become so attractive for researchers. As a social media platform, Twitter is a relatively small platform compared to Facebook. It only has about 300 million monthly active users, whereas Facebook has over 2 billion monthly active users.[19] Twitter's minor position on the social media market, and the more specified user group are recurring point of critiques to use the platform for

---

[16] Halavais, 'Structure of Twitter', p. 36.
[17] K. Grifantini, 'The Evolution of Retweeting', *Technology Review,* 26 August, 2009 <https://www.technologyreview.com/s/415043/the-evolution-of-retweeting/> (18 September, 2017)
[18] Halavais, 'Structure of Twitter', 30-31.
[19] Zephoria, 'Strategic Insights' <https://zephoria.com/top-15-valuable-facebook-statistics/>(18 September, 2018).

research.[20] A quick comparison of the structures of the two platforms clarifies why Twitter is considered to be more straightforward than Facebook, however. On Facebook users have mutual friendships, but there is also an option to follow or get followed. In addition, they can like pages that can represent celebrities, companies, political parties, music groups or simply a daily doses of memes, among other phenomena. Users can unite in groups that are either open, closed or completely secret. As a result, relations are far more complex. Topical discussions often take place on multiple pages and groups all over the platform. As each user has connection with news sources of their preference, discussions arise all over the platform. There is a hashtag option inspired by Twitter's operator, but is only used scarcely. Therefore it is much harder to capture a discussion around a certain topic on Facebook. It might be

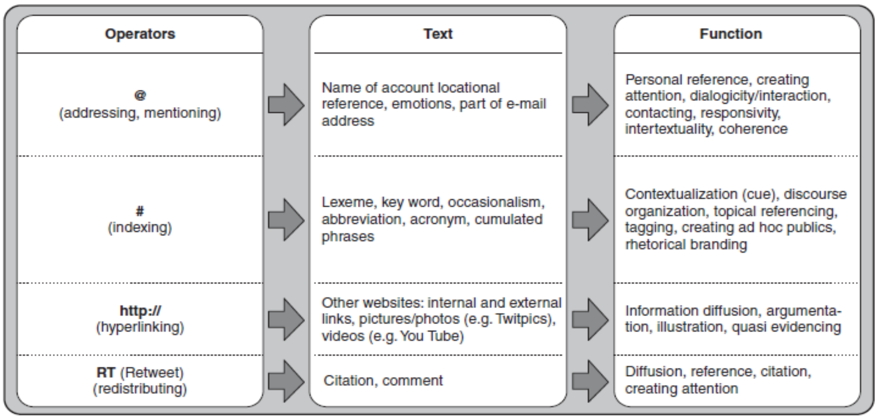


Figure 2: Functional operator of Twitter by Thimm, Dang-Ahn and Einspänner. Source: *Mediatized worlds* (2014).

richer in its content and in the diversity of its users, but its textual content is harder to filter.

Twitter's clear structure, by contrast, has allowed scholars to capture the main functions of the operators in models. The four operators and their typical text forms and functions have been brought together in a functional operator model. Models like this, and the structural communication model, provide other scholars with a clear view of Twitter's structure, from where they can start thinking of and shaping their research questions. Along with the developments of Twitter's four operators, new types of studies emerged. The next paragraph will explain how types of research developed parallel to the developments of Twitter's structure.

[20] J. Ruiz Soler 'Twitter Research for Social Scientists: A Brief Introduction to the Benefits, Limitations and Tools for Analysing Twitter Data.' *Dígitos: Revista De Comunicación Digital*, no. 3 (2017), pp. 17-32.

1.2 Using Twitter's Database for Research

In *Twitter and Society* (2013) Twitter's development as a platform suitable for scholarly research is described in three phases; Twitter I, II & III. The three phases represent the evolution of types of research that followed from the platform's development. Twitter I (2006 – 2009), the first generation of Twitter, was an urban lifestyle tool for friends to provide each other with updates about their whereabouts and activities. The central question asked by Twitter above the status bar was 'What are you doing?'. Within in this phase the majority of the tweets remained unstructured, which resulted in studies that categorized tweets, to indicate whether they are suitable for research. The second phase (Twitter II, 2009 – 2011) is characterized by the change of this question in 2009 from 'What are you doing' to "What's happening'. This change shifted the focus from personal to topical activity, making it a backchannel to follow and discuss ongoing events with the help of hashtags. Within these hashtags structured conversation arose, allowing researchers to study the development of events and discussion on social and political topics. The last and current phase, Twitter III (2011-), approaches the platform as an archived data set, where both public and commercial institutions attempt to preserve all Twitter data, to allow historical research of tweets.

Twitter was founded in 2006 by Jack Dorsey along with co-founders Evan Williams and Biz Stone. It responded to new media trend of dispatching short messages, that was at that point gaining popularity trough SMS messages. As mentioned in the introduction, the 140 character length of a tweet is derived from SMS. The focus on telling where you are and what you're doing was inspired by other systems. In an interview with the *Los Angeles Times* Dorsey tells 'Twitter has his conceptual roots in the world of vehicle dispatch – where cars and bikes zooming around town must constantly squawk to each other about where they are and what they're up to.'[21] He applied these systems to the social, mobile Web, so anyone could squawk form anywhere. This is still the Big Idea behind the platform as their current mission is to 'offer everyone the ability to create and share information immediately, without any limitations.'[22]

Dorsey had put his ideas on paper in 2000, and the first sketch is still present in his

[21] D. Sarno, 'Twitter creator Jack Dorsey illuminates the site's founding document. Part I', *Los Angeles Times*, 18 February, 2009 <http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html>(18 September, 2017).

[22] Twitter, 'Company#about' <https://about.twitter.com/nl/company#about>(18 September, 2017)

office and can be found online.[23] He put his idea into practice by writing a small program for the RIM 850, the predecessor of the BlackBerry, that was basically a small email device. It allowed him to write e-mails consisting of four lines and share his whereabouts and activities, to a list of his friends and allowed them to reply theirs. It worked fine, but it had the limitation that apart from Dorsey and his friends the devices were hardly used by anyone.[24] A technical leap forward that made mobile devices affordable, comfortable and interoperable was necessary to launch a platform suitable for mass use. By that time the web was already evolving to the Web 2.0, on which users are responsible for the generated content and its distribution. Yet, Twitter's success depended on multiple technological developments. The growing interactivity on the Web 2.0 was a good start for the first years to share 'what you were doing', but to share 'what's happening', at any time and from anywhere, other types of technologies besides the Web 2.0 were necessary. It demanded mobile connectivity, faster mobile internet and better cameras to make it into the continuous stream of information that it is today. Therefore, Twitter I is marked by users who were mostly sharing their whereabouts with their friend/follower network, rather than participating in public discussions or reporting news from the ground.

Twitter's first phase as an ambient and friend-following medium inspired research that focused on the categorization of the content of the tweets, addressing the question whether most of its content was banal or not. BBC news categorized 2.000 tweets and categorized these using the labels 'pointless babble', 'conversational', 'pass-along value', 'self-promotional' and 'spam'. Their intention was to study the platform for its potential as an information source. Only tweets categorized as pass-along value were considered to be informational, a category that was good for 8,7% of the total tweets analyzed. The main finding, however, is indicated in the title of the article: 'Twitter tweets are 40% babble'. The label 'babble' was assigned to tweets of the 'I'm eating a sandwich' type.[25] Most of the initial Twitter studies had a similar approach. They focused on the question whether Twitter was a usable source for information at all. Categorizing tweets was a popular way to do so. Next to 'pointless babble', tweets in the category 'daily chatter' offered opportunities to study communication on Twitter. With the help of user innovation, the '@' symbol became the common symbol to reply to particular users, making it more easy to define communicative

[23] The photo is also added to Sarno, 'Twitter creator Jack Dorsey. Part I'.
[24] Sarno, 'Twitter creator Jack Dorsey'.
[25] Anon., 'Twitter tweets are 40% babble', *BBC News,* 17 August, 2009 <http://news.bbc.co.uk/1/hi/technology/8204842.stm>(18 September 2017).

tweets. The hashtag (#) was also a product of user innovation, first used to report about the San Diego fires in 2007 (#sandiego), making it easy to classify tweets that are 'reporting news'. In the same way Twitter users started sharing URLs, which were initially taking up a lot of the 140 characters, but Twitter quickly saw the potential and offered the opportunity for shortened URL's. This generated the third user innovated category of 'sharing information'. Although the help of user innovation simplified categorization for researchers, personal tweets remained dominant. Until 2009 up to 80% of the tweets consisted of personal information. Therefore, Twitter I is defined as a period 'inconsequential information'. During Twitter I structured use of the platform, for example to report on news events with hashtags, was still in development and not the norm.[26]

. This changed when Twitter's tagline was changed from 'What are you doing' to 'What's happening' in 2009. The platform made a move from 'an ego to a reporting machine'.[27] This change was accompanied by the introduction of a trending topic feature in April 2009. Twitter co-founder Biz Stone described the new purpose of Twitter as becoming a 'state of affairs machine', or 'discovery engine for finding out what is happening right now'.[28] Dorsey stated that he indeed noted that Twitter did 'well at: natural disasters, man-made disasters, events, conferences, presidential elections', or what he calls 'massively shared experiences'.[29] Changing the tagline was a move inspired by users who discovered Twitter's possibility to be used as a backchannel. The founders of Twitter picked up and improved methods that were being innovated by users. In 2007 and 2008 users already utilized the platform to comment on speakers on conferences. This resulted in standardized hashtags for conferences. In the same manner breaking events such as the San Diego fires in 2007, the Sichuan earthquake in May 2008, the Mumbai terrorist attacks in November 2008 and James Karl Buck's arrest in Egypt in 2008 were covered and commented on by Twitter users with the help of hashtags. When US airways flight 1549 crashed into new York's Hudson river,

[26] R. Rogers, 'Debanalising Twitter: The Transformation of an Object of Study', in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014), pp. ix-xxiii, p. xiii.
[27] R. Tate, 'Twitter's new prompt: A linguist weighs in', *Gawker,* 19 November, 2009 <http://gawker.com/5408768/twitters-new-prompt-a-linguist-weighs-in> (18 September, 2017).
[28] B. Stone, 'Twitter search for everyone!', *Twitter Blog,* 30 April, 2009 <https://blog.twitter.com/official/en_us/a/2009/twitter-search-for-everyone.html> (18 September, 2017).
[29] D. Sarno, 'Jack Dorsey on the Twitter ecosystem, journalism and how to reduce reply spam. Part II', *Los Angeles Times,* 19 February, 2017 <http://latimesblogs.latimes.com/technology/2009/02/jack-dorsey-on.html>(18 September, 2017).

Twitter users broke the news with eyewitness reports including photos from the scene.[30] Twitter coverage of these events are often listed in early Twitter studies.[31]

The changed tagline, trending topic feature and the standardized use of hashtags expanded Twitter's function as a news medium for event following, sparking hashtag-based studies that define Twitter II. One of the challenges created by Twitter's uprising function as a news medium was whether the platform could be 'made into a storytelling machine that recounts the events on the ground and on Twitter?'.[32] In the run-up to upcoming presidential elections in Iran, on 12 June 2009, Twitter was being watched carefully by political bloggers and tech watchers**.** The results of the election contained many irregularities, leading to demonstrations by the Iranian Green Movement. The American political bloggers Andrew Sullivan and Ari Berman released articles titled 'The Revolution Will Be Twittered' and 'Iran's Twitter Revolution' on 13 and 15 June, when the protest where at its heaviest. They considered Twitter to be a revolutionary technology, that was in lineage with the fax machine, the mobile phone and text messaging. Twitter's potential revolutionary use lead to a debate between Clay Shirky and Evgeny Morozov, who both study the social and economic impact of internet technologies. Shirky's book *Here Comes Everybody* (2008) depicted social media as a democratizing force, that could change the course of history. Morozov debunked this idea, especially for Twitter, using the demonstration in Iran as an example. He argued that the great majority of the content was generated by a very small number of people. Iran's Twitter users during the protests were mostly 'Pro-Western, technology-friendly and iPod-carrying young people' representing only a 'tiny and, most important, extremely untypical segment of the Iranian population'. Despite the extensive use of Twitter by Iranian opposition leader Mir-Houssein Mousavi, he could not utilize the platform for a political breach with the revolts. This was partly due to the shutdown of the Internet and mobile network by the authoritarian government. The Iranian government survived, and issued counter measures resulting in the persecution of critical bloggers, journalists and a series of violent crackdowns on politically active university students. [33]

In the aftermath of the Iran's election, Twitter's function as a tool for reporting on the ground, had attracted scholars in two ways. Firstly, the election gave rise to a debate on

---

[30] C. Beaumont, 'New York plane crash: Twitter: breaks the news, again.' *The Telegraph,* 16 January, 2009 <http://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html> (18 September, 2017).
[31] R. Rogers, 'Debanalising Twitter', p. xvii.
[32] Ibidem, p. xx.
[33] Ibidem, p. xvii-xix.

whether Twitter was a revolutionary internet technology. Secondly, the tweets posted with the hashtag #IranElection were gathered to create a suitable method to create the narrative of the Iranian protest on the basis of the reports from the ground. Researchers have experimented with these two new possibilities using the tweets connected to the data set of #IranElection, and other related hashtags. In contrast to Twitter I, these tweets were already categorized under the hashtags, and provided information about broader societal developments, rather than on personal matters of individuals. The overview and chronology of the conversation, size and nature of the user population, as well as the role of prominent and influential users within this population, became important factors in order to gain insight for both issues. The results of *The Iranian Election on Twitter: The First Eighteen Days* showed that indeed only a fraction of the Iranian population tweeted actively during the protest (480.000 users on 70.000.000 inhabitants) and an even smaller community (top 10 % of users) accounted for 65,5% of the tweets. One out of four tweets was a retweet. With the help of the top three retweets per day, the study *For the ppl of Iran --#iranelection RT* was able recreate the narrative of the riots following the election.[34] So the Tweets proved to be sufficient to 'recount the event on the grounds'. Yet, it is important to note that the tweets just represent *a* narrative, which is the narrative of an 'extremely untypical, young, technology friendly' group of Iranians as Morozov stated.

With the help of the data derived from hundreds of thousands of short messages, Twitter's revolutionary potential was debunked, as the results showed that only a small unrepresentative section of the Iranian population used Twitter, whom furthermore were disarmed of their ability to tweet, by shutting down the internet.[35] Yet, its potential as an informational platform could clearly be demonstrated. As Twitter's informational potential was growing, so did the number of users and tweets sent on a daily basis. During the phase of Twitter II (2009-2011), the platform grew exponentially. According to Twitter's own data, users on Twitter were sending up to 65 million tweet in June 2010, while in January 2009 only 2 million tweets were sent on a daily basis. It eventually grew to a total number of 200 million tweets in 2011.[36] The exponential growth continued until 2015 when more than 500 million tweets a day were being sent. Currently the platform is still growing, but at a slower pace. Worldwide the number of monthly active users has grown from 302 million in the first

[34] R. Rogers, 'Debanalising Twitter', p xix-xxi.
[35] Ibidem, p. xviii-xix.
[36] Twitter Engineering, '200 million Tweets per day', *Twitter Blog,* 30 June, 2011 <https://blog.twitter.com/official/en_us/a/2011/200-million-tweets-per-day.html>(18 September, 2017).

quarter of 2015, to 328 million in the second quarter of 2017.[37] The current number of daily tweets is debated though, as third party data tools, see a decline, but Twitter denies this without any further explanation.[38]

Twitter's growth and the technological trends that stimulate mobile connectivity turned the platform into an attractive data set, suitable for interdisciplinary use. Tweets can be retrieved trough the platform's API, commercial data collectors and (online) software programs. The exact possibilities, differences and limitations between these methods, will be explained in the next chapter. Yet, retrieving and storing data from Twitter is considered to be relatively easy. In adjustment, the inbuilt tools like the retweet, @replies, follower-followees network and categorization by hashtags, provide a clear starting point for researchers. Twitter III (2011-) defines Twitter as an archived data set, allowing researches to track tweets back in history. However, *Twitter & Society* was published in 2014, four years after the Library of Congress in the U.S. announced they would archive every public tweet since Twitter's inception in March 2006. The prospect was that scholars could access every tweet that was at least six month old. With this commitment the library recognized the historical, informational and cultural value of Twitter data for future research. Until now, however, they have not been able to realize their plans and the archive is still unavailable, leaving hundreds of inquiries by researchers unanswered. Meanwhile commercial data institutions like Gnip fill the vacuum, by separately selling historical tweets from Twitter's inception in March 2006 and onwards. Therefore the prospect of Twitter III, as openly accessible data source for research, has not been fulfilled yet.

1.3 The Crossroad of the Humanities and Twitter

Twitter is an outstanding platform for humans to express their ideas, opinions, imagination and experiences. Therefore, it is easy to link it to the humanities, the study 'of how people process and document the human experience'.[39] The humanities have been defined in many ways. There are characteristics that recur throughout all the definitions, nevertheless. The disciplines that generally fall under the humanities umbrella are language, literature, history, jurisprudence, philosophy, comparative religion, ethic and the arts. These subjects are

[37] Twitter Files, 'Selected Company Metrics and Financials', <http://files.shareholder.com/downloads/AMDA-2F526X/5225764351x0x951002/FCE28680-E74E-4349-A11C-4B86BBABFB26/Q217_Selected_Company_Metrics_and_Financials.pd>(10 October, 2017).

[38] J. Edwards, 'Leaked Twitter API data shows the number of tweets is in serious decline', *Business Insider UK,* 2 February, 2016 <http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2?international=true&r=UK&IR=T>(10 October, 2017).

[39] Stanford Humanities Center, 'What are the Humanities?' <http://shc.stanford.edu/what-are-the-humanities> (18 September 2017).

considered to be ‘modes of expression’. The humanities use our (documented) memory and imagination to describe ‘where we have been and helping us envision where we are going’ and ‘how people have created their world, and how they will they in turn will be created by it’. The methods used by humanists ‘are primarily critical, or speculative and have a significant historical element’. Thereby, the humanities distinguish themselves from the natural and most social studies with interpretive research in contrast to empirical research.[40]

Twitter as a research platform can be considered as a big data platform. Computational methods, often empowered by empirical statistics and algorithms, are needed to analyze the vast number of messages and the 160 pieces of metadata that come with each message. Given such exigencies, the use of Twitter data seems less probable for humanities scholars. Yet, the humanities have never been completely obsolete of data. Demographics, climate change and civil administration are all factors that can help to shape a historic narrative. Numbers were therefore helpful to narrate human (contemporary) history and experience. Halfway the 20th century humanities scholars started to experiment with the reverse process. They started to turn texts into numbers. FR Roberto Busa S.J. (1913 – 2011) is considered to be the founding father of this process, which is now commonly known as the Digital Humanities. In collaboration with IBM, he started to work on an index for the complete works of Thomas Aquinas. This consisted out of 1.5 million lines and 9 million medieval Latin words. The goal was to use algorithmic processes in order to make the corpus searchable for terms, word counts and word concordance. To make this possible, the text had to be broken down into separate phrases and pressed on punch cards. Each sentence card had to be multiplied as many times as there were words on each. [41] The biggest challenge was lemmatization, the act of grouping together different inflected forms of a word so they can be analyzed as a single item. It was his life’s work, that resulted in a printed form in 1974, a CD-rom version in 1992, and a web based version in 2005. As word based searches define daily internet use nowadays, Busa is not only seen as the pioneer for the digital humanities, but also as a forerunner of combining informatics and the written word, two fields that have become inseparable from the digital world, in which we daily surf the internet or write emails.[42] Even Twitter’s ability

Peter Verhaar 29-11-17 21:22
Deleted:

[40] The quotes are taken from a variety of definitions given on 4Humanities, ‘What are the Humanities’ <http://4humanities.org/2014/12/what-are-the-humanities/>(10 October, 2017).

[41] J. Norman. ‘Roberto Busa & IBM adapt Punched Card Tabulating Sort Words in a Literary Text: The Origins of Humanities Computing’, *HistoryofInformation.com* <http://www.historyofinformation.com/expanded.php?id=2321>(18 September, 2017).

[42] E. Priego, ‘Father Roberto Busa: One academic’s impact on HE and my career’, *The Guardian,* 12 August, 2011 <https://www.theguardian.com/higher-education-network/blog/2011/aug/12/father-roberto-busa-academic-impact>(18 September, 2017).

to discover trending topics and/or terms, can be viewed as a modern example of Busa's initial idea.

Although Busa's idea was revolutionary, the punch card technology was expensive, time consuming and the process was hard to control. During the first decades of the digital humanities, or humanities computing as it was named from 1949 until about 2002, researchers were dependent on computer labs. This resulted in a rather small community of humanities scholars who could only work in the field if they had sufficient knowledge of programming. This lead to little methodological development during the 1960s and 1970s. Counting words was one of the most common methods, mostly with the goal to calculate the vocabulary complexity of texts. This way scholars hoped to identify the authorship of for example Shakespeare, by comparing the presence words with two, three of four letters.[43] Work in this early period was hampered by the technology of batch processing systems, of which the punch card system was an example. Yet, the field also organized itself as the first conferences were being held with a series of six conferences in 1964 and 1965 organized by IBM. After 1972, conferences became regular occurrences, and the first associations were founded in 1973 and 1978. Yet, the journals and conferences mostly reached those involved with humanities computing, and did not lead to many publications in traditional humanities journals.[44]

Humanities computing was able to make a leap forward in the 80s with the help of two new technologies: the personal computer and electronic mail. The personal computer gave researchers the freedom to experiment without the consent of computer centers and electronic mail allowed them to debate and review each other's work. In 1987 the first electronic seminar was hosted, which to this day remains an active and important venue for digital humanities researchers. In the same year the Text Coding Initiative (TEI) was founded, which set guidelines for making digital texts machine readable. Projects like TEI standardized the methods to prepare texts for computational analysis, making it more understandable and accessible for a larger group of humanists. In addition to electronic mail, the internet, which became a major means of communication in the course of the 1990, played a large role in de

---

[43] T. C., Mendenhall, 'A Mechanical Solution of a Literary Problem', *The Popular Science Monthly* 60: 97–105 (1901).

[44]S. Hockey, 'The History of Humanities Computing' in S. Schreibman, R. Siemens and J. Unsworth, *A Companion to Digital Humanities* retrieved from <http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1&toc.depth=1&toc.id=ss1-2-1&brand=default>(18 September, 2017).

developments of the digital humanities, as it brought new opportunities for the publication and the dissemination of digital projects.[45]

Almost sixty years after Busa started to transform texts intro numbers, the biggest change to the digital humanities is its accessibility. It started as a small community of humanists with extensive technical computing knowledge, but nowadays preprogrammed software and structured rules and methods make computer-based methods available for humanists with no programming knowledge at all. Anyone with access to a computer, can now analyze Shakespeare with a few clicks.[46] This allows a growing group of humanists to answer different questions, or to adapt traditional ones. Computational analysis, for example, is a useful tool to study vocabulary change through the years with the help of digitized dictionaries. This has two benefits. Firstly, it saves much time in contrast to manually comparing dictionaries. Secondly, it can reveal structural changes. As is the case for Twitter, such new online platforms highlight the necessity of computational methods to answer research questions, because the number of tweets can simply be too large to read manually. Although information overload has been a problem for centuries, digital social media platforms make it possible to experience it live. At the most popular hashtags dozens of new tweets can be posted in the time you can only read a couple of them. As computer scientist and philosopher Jaron Lanier stated 'It's as if you kneel to plant the seed of a tree and it grows so fast that it swallows your whole town before you can even rise to your feet'.[47]

The digital humanities especially work computationally with tools to handle these massive quantities of text, and to extract data from these. This process is called text mining or textual analysis and it can be applied to large textual datasets like archives of books, newspapers, journal articles but also social media messages. Apart from simple word counts, these tools can be used to preform sentiment analysis. Sentiment analysis can differentiate subjective forms of textual expression, for example political opinions and emotional responses. In chapter 3 some studies will be evaluated that order these modes of expression, to locate different collective identities or political statements in tweets. Consequently, these tools can be used to transform the data into visualizations in the form of line diagrams or cluster networks. Both chapter 3 and 4 will demonstrate how visualizations are made and read.

---

[45] S. Hockey, 'The History of Humanities Computing'.
[46] For example on this site: Voyant Tools <https://voyant-tools.org/>(18 September, 2017).
[47] J. Lanier, *You Are Not A Gadget: A Manifesto* (London: Penguin, 2010).

Thus, Twitter and the humanities cross on two sections. First and foremost, the platform documents human expression, suitable to analyze human society and culture. Secondly, tools used by the digital humanities can be utilized to analyze Twitter. In these thesis Twitter studies will be divided into two categories to analyze their potential for the humanities. The first category will consist of research that is primarily focused on the metadata around tweets. This mainly includes research considering social and scholarly communication networks and geolocation. The second category will focus on research that primarily focuses on the content of the text in the tweets, by analyzing political statements, collective identities, discourse, personalities and historical commemoration.

Chapter 2: How do we study Twitter, what are the limits?

Despite improved technological insight and development there are still challenges to overcome for humanities scholars before they can productively exploit Twitter as a research platform. Both policy-oriented and technological challenges have to be overcome in order to retrieve data for research. These challenges and how they can, or cannot be overcome, will be explained in the following chapter.

2.1 Twitter's API: Technical and Policy Challenges

Twitter produces millions of messages a day that can contain up to 160 pieces of metadata. Some of the most useful metadata among these 160 pieces are the location, language, time of creation and number of retweets of a tweet but also user information like a name, time zone of the used computer or mobile, date and time of account creation, follower count and the personal biography.[48] Therefore, collecting, archiving and filtering a dataset are the first and most complicated challenges for researchers. How do they get the data from Twitter's database on their personal computer? This is done with Twitter's free Application Programming Interface (API). An API is the messenger that takes requests and tells a system what you want to do and then returns the response back to you. The API can be compared with a librarian, where researchers can request a specified series of documents, for example all local newspapers in Amsterdam from 1980 until 1990, which can then be delivered by the librarian. However, Twitter's API offers more opportunities than requests only. Users can build third-party applications to automatically collect data, but also to integrate video's, photo's, music and links from other platforms into tweets. Twitter's open API combined with its simple structure lead to improvements and adaptation by its users, like the shortened URLs, inclusion of YouTube videos or automatically sharing your tweets on Facebook.

Despite Twitter's open API, users cannot freely request all tweets from Twitter's first years, or any certain historical period. This is partly due to the technical limitations, but mostly due to the commercialization of data. The main producers of Twitter's data, its users, have the least control over it. Those who create the data no longer have free access to the complete database. Data is considered to be a 'commodity on a par with scarce natural resources', and is therefore only shared for a price by Twitter.[49] Social media data has attracted companies that collect and sell the data. One of these companies is Gnip. They do

[48] W. Wolny, 'Knowledge Gained from Twitter Data', *Annals of Computer Science and Information Systems,* 8, pp. 1133-1336.
[49] C. Puschmann and Jean Burgess, 'The Politics of Twitter Data' in Weller et. al (ed.), *Twitter and Society*, pp. 43-54.

provide the full archive of Twitter data, enabling researchers to find and analyze any public Tweet posted since Twitter's origin in 2006.[50] Yet, this comes at a price of thousands of dollars a month, depending on the service and quantity researchers are after.[51] Currently Gnip is owned by Twitter, as they acquired it in April 2014 for $134.1 million.[52] As a platform Twitter combines communication and publication. It has a very broad copyright license on its users' content, as its terms of service shows;

> 'By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same.'[53]

Twitter's corporate approach is part of larger shift from an 'open' Internet to a more 'closed' internet. On the open internet users could participate and innovate trough Twitter's API and thereby improve and adapt the platform to its current form. On the closed internet users are only the suppliers of the main product, that is data, which is traded between a few corporate superpowers that set the rules, and have the capital to buy other upcoming platforms and thereby monopolize the Internet's main communication platforms.[54] This process is often called gatekeeping, the process where in this case a strong media company like Twitter decides which content (or data) is, and which is not available for the public. Following this trend, Twitter has deliberately reduced the openness of its API to third-party applications. From 2011 onwards their policy became more restrictive, denying applications to access and store quantities of data. Following this trend, Twitter has either bought or cut off other applications.[55]

---

[50]The historical API can be accessed here: Gnip, 'Historical' <https://gnip.com/historical/>(18 September, 2017).
[51] In 2010, access to 50% of Tweets from Gnup cost about $30.000 and acces to 10% cost about $5.000 per month. D. Gaffney and C. Puschmann, 'Data Collection on Twitter' in Weller et. al (ed.), *Twitter and Society*, pp. 55-68.
[52] Y. Koh, 'Twitter Paid $134 Million for Data Partner Gnip', *The Wall Street Journal,* 11 August, 2014 <https://blogs.wsj.com/digits/2014/08/11/twitter-paid-134-million-for-data-partner-gnip/> (18 September 2018).
[53] Twitter, 'Terms of Service' <https://twitter.com/en/tos>(18 September, 2017).
[54] Puschmann, 'The Politics of Twitter Data', p. 45.
[55] J. Valentino-De Vries, 'Twitter Buys TweetDeck', *The Wall Street Journal,* 25 May, 2011 <https://blogs.wsj.com/digits/2011/05/25/twitter-buys-tweetdeck/> (18 September 2018) and C. Warren, 'Twitter's API Update Cuts Off Oxygen to Third-Party Clients', *Mashable,* 16 August, 2012. <http://mashable.com/2012/08/16/twitter-api-big-changes/#S2I6Tlp74uq2>(18 September, 2017).

There are, however, still free and open opportunities for those who wish to study Twitter with the help of its API. Twitter offers two different data interfaces for researchers: the Streaming API and the REST API. The Streaming API is the most widely used API for large-scale quantitative analyses of Twitter data. It is a somewhat complicated method though, because it does not request data from the database itself, but collects data which is at the moment of request generated live by Twitter users. It collects data with a 'push' based strategy instead of a 'pull' based strategy. Once a request for data retrieval is made to the Streaming API it provides a continuous stream of public information from Twitter. It is a live polling system, not suitable for a historical analysis that needs a 'pull' system that retrieves tweets from a specified period. One of the complications for researchers with the streaming API is the limitation of its use for scheduled events, like elections, debates, TV-shows and sport events. It cannot be used effectively in hindsight though, as it will miss the beginning and origin of an event, and is therefore not ideal for unpredicted events like natural disasters and terroristic attacks.[56]

Considering Twitter's corporate approach, collecting 100% of the tweets free of charge via Streaming API is not possible. In fact, the Streaming API has three different bandwidths: 'spritzer', 'gardenhose' and 'firehose', which respectively deliver up to 1%, 10% and 100% of all tweets posted at the system. As companies like Gnip offer the opportunity to retrieve all tweets from a given moment, logically the firehose bandwidth is only available for those who have a business relationship with either the commercial data company or Twitter itself. Gardenhose on its turn, is granted occasionally to users with compelling and defensible reasons for increased access. Only the spritzer bandwidth is freely available for everyone, as long as they have a Twitter account. The 1% bandwidth is sufficient to collect tweets around small events and congresses or in languages that are less represented on the platform. Yet, when the tweets concerning an event exceed the 1% percent of all tweets on Twitter, the results are sampled.[57]

Focusing on a particular event can be done with the help of request parameters. Parameters help to filter the requested data. Besides filtered data, one can also choose to get a *sample*. This method offers the opportunity to get 1% or 10% (depending of the acquired bandwidth) of all tweets at random. Opposite to this sample function, the Streaming API has a

---

[56] G. Puschmann, 'Data Collection on Twitter'.
[57] Ibidem.

filter function, which is divided into four more parameters: *language, follow, track* and *location*. The have the following function:

- *Language*: detects tweets written in a specific language, in order to retrieve only tweets written in this language
- *Follow*: returns tweets from a set of users represented by their collective comma-delimited user IDs. This will include al tweets, retweets by the user and of the user's tweets, manual replies and replies to tweets. It excludes any tweets with mentioning, manual retweets and tweets from protected users.
- *Track*: allows users to create a comma-separated list of phrases that returns only tweets including the words separated by non-word characters. This is pre-eminently the method to retrieve Tweets containing a certain hashtag or commenting on a certain event or person. Basically it works like most advanced search tools, where terms can be combined (AND) or separated (OR).
- *Locations*: provides the opportunity to collect Tweets with the help of comma-separated longitude and latitude values. This will return only Tweets that are 'geotagged'. These Tweets are either represented as points when an exact location is retrieved, or as rectangles of four pairs of points that can be as small as a city park or as large as a province. The most recent studies that tried to determine the percentage of geotagged Tweets estimated it was approximately 1%-2%. These studies are at the moment of writing more than four years old.[58] As smartphone use increases and becomes more advanced the percentage is probably higher now, but this has not yet been researched by new studies.[59]

These parameters allow users to focus for example on a particular hashtag, used to comment and inform on an event. But what if a researchers has missed the first half hour of an unscheduled event? Or what if researchers want to recreate a spontaneous demonstration from the day before? To what extend is it possible to retrieve this half our or day of missed tweets? This is where the REST (REpresentational State Transfer) API comes into use. The REST API uses a pull strategy for data retrieval, meaning users can explicitly request Tweets from the database. Tweets that were missed with the Streaming API, can be recovered with the REST API, yet this also has its limits.

---

[58] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, A. & E. Shook, 'Mapping the global Twitter heartbeat: The geography of Twitter', *First Monday, 18*(5) (2013) <http://firstmonday.org/article/view/4366/3654>(28 September 2017).
[59] G. Puschmann, 'Data Collection on Twitter'.

The Search API is part of the REST API, and allows users to retrieve defined phrases, in the way the *track* function of the Streaming API can be used to collect all Tweets with a certain hashtag. The first and most important limit is that this can only be done for Tweets that are no more than a week old. According to Twitter the Search feature focuses 'on relevance and not completeness'.[60] Therefore, the second limit is that the results are sampled. The third limit is the rate limit. Users can only send a maximum of 180 requests every 15 minutes. With each request, the Search API can deliver up to 100 tweets. So a single users can retrieve up to 72,000 tweets per hour and a total of 1,728,000 tweets per day, considering a user sends outs request for a 24-hour cycle.[61] This is a fraction of the 300 million tweets sent on a daily basis. Whiles the Streaming API is the preferred methods to capture tweets itself, REST API proves is value for more static data and therefore to create lists of followers from a user, its entire tweeting, retweeting or favoring history.

2.2 Is a data sample sufficient?

Both the use of the Streaming and the REST API might lead to a situation were merely a sample of all targeted tweets can be collected. Although Twitter offers an elaborate guide to its API, the company does not share its sampling method. This issue inspired some researchers to statistically analyze Twitter's sampling methods, in order to answer the question whether sampled datasets give a sufficient representation of the activity on Twitter as a whole. One of the first evaluations of sampled data was done by Morstatter et. al, who compared the results from the free Streaming API granting 1% of all data, with a corresponding 100% firehose feed provided by Twitter. They used both methods to discover top hashtags and for topic analysis. Tweets where gathered with the help of parameters that aimed for specified hashtags within the geographical boundaries of Syria. The results revealed some limitations of the Streaming API compared with firehose feed. Firstly, the 1% bandwidth resulted in less coverage when the number of tweets within the set of parameters grew. This indicated that even when data retrieval is specified by a query of parameters, the 1% bandwidth can be insufficient. Secondly, the top hashtags were retrieved more adequately

[60] Twitter Development Documentation, 'The Search API' <https://dev.twitter.com/rest/public/search>(18 September, 2017).
[61] Twitter Development Documentation, 'Rate Limits: Chart' <https://dev.twitter.com/rest/public/rate-limits> (18 September, 2017).

for a large number tweets, but were often misleading for a small number of tweets. Thirdly, topical analysis became more accurate as the Streaming API delivered more data. [62]

These results could all be explained simply by the fact that the free Streaming API delivers only up to 1% of all Twitter data. Therefore Morstatter et. al also compared the results from the Streaming API with 100 random samples taken from the Firehose feed. The random samples from the firehose feed performed better than the Streaming API. The random samples had positive correlation with the complete firehose feed considering top hashtags and topical analysis, while the Streaming API had a negative correlation. The Streaming API only showed an equal performance with the geotagged tweets. Here it delivered the complete set of geotagged tweets, this is probably due to the already low percentage (1%) of geo-tagged tweets.[63] These results show that the Streaming API follows a set of rules, that leads to a biased data set, something researches have to take into consideration when they stick to the free Streaming API and aim to collect data exceeding the 1% bandwidth.

In another study Wang et. al compared samples from the Streaming's API spritzer (1%) and gardenhose (10%) with a complete corresponding Twitter dataset collected with REST API in a somewhat different manner. Unlike Morstatter they did not collect their tweets around event based specified hashtags, but collected the complete set of tweets from the Singaporean Twitter users during May 2012. By comparing the samples with the complete numbers, Wang discovered that they return 0,95% (spritzer) and 9,6% (gardenhose) of the total tweets on average. The samples proved themselves to be adequate concerning their representation of the users daily activity. The only small connotation to be made of the subject of user activity, is the tendency of small samples (spritzer) to overestimate the role of low frequency users. The spritzer sampling ratio proved to be sufficient though to capture important tweet content like text terms and URL domains and the frequency of appearance of the content terms. In adjustment, Wang states that sampled datasets are viable to use for 'tasks such as event detection, sentiment analysis and tweet summarization'. The small spritzer sample was not suited to provide adequate information concerning hashtags. According to Wang, the larger gardenhose sample is needed for this. Another trend Wang's research exposed, is that Twitter's samples focus on its representative users. Users who post less than

---

[62] F. Morstatteret al., 'Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose', (2013).
[63] F. Morstatteret al., 'Is the Sample Good Enough?'.

one tweet on a daily basis often remain unsampled. Their opposites, users who are highly active, but are probably spammer bots, are deliberately left out of the samples by Twitter.[64]

2.3 The Library of Congress Twitter Archive

Exploring Twitter's API and its policy made clear that scholars need technological know-how about Twitter's API, insight into the validity of samples, money and computing power when a large (historic) dataset is desired. The opportunities the Twitter data sets offer, are therefore not easy to seize for scholars. Especially for humanities scholars this can be a reason to avoid Twitter research, because the effort they have to put in might not meet their expectations. Yet, there is still hope for accessible free and tailored Twitter data, as Twitter agreed to donate every public tweet since its inception in March 2006 to the Library of Congress. The so called 'gift agreement' was made in 2010 and would be very helpful for researchers to overcome some technical barriers.[65] However, until this day of writing the archive is still unavailable. So far only a pilot project called Twitter Data Grants has allowed researchers to obtain free data access to Twitter datasets.[66] Despite their intentions, Twitter only granted free access to only six of the in total over 1,300 research proposals. [67]

The very limited access granted by Twitter has urged the necessity of an open Twitter archive for scholarly research. What are the challenges faced by the Library of Congress that prevented them to realize this for the past seven years? With 500 million tweets sent on a daily basis, the Twitter database provides a continuous stream of information, or a some call it, a 'flood of information'.[68] The Library of Congress is not unfamiliar with huge quantities of data though. Holding more than 36 million books and printed materials, as well as more than 121 million maps, manuscripts, photographs, films, audio and video recordings, prints, drawings and other special collections, it is the largest library in the world. They are also familiar with conserving digital data, as they have been operating a web archiving program since 2000. In 2014 they had collected up to 525 terabytes of Web archive data. They especially aimed at sites concerned with national issues like , U.S politics and national

---

[64] Y. Wang, J. Callan & B. Zheng, 'Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API' *ACM Transactions on the Web (TWEB)*, 9(3), (2015), pp.1–23.
[65] M. Raymond, 'Twitter Donates Entire Tweet Archive to Library of Congress', *Library of Congress*, 15 April, 2010 <https://www.loc.gov/item/prn-10-081/>(18 September, 2017).
[66] R. Krikorian, 'Introducing Twitter Data Grants', *Twitter Blog*, 5 February 2014 <https://blog.twitter.com/engineering/en_us/a/2014/introducing-twitter-data-grants.html>(18 September, 2017).
[67] R. Krikorian, 'Twitter #DataGrants selections', *Twitter Blog*, 17 April, 2014 <https://blog.twitter.com/engineering/en_us/a/2014/twitter-datagrants-selections.html> (18 september, 2017).
[68] M. Ingram, ''Drinking from the Twitter firehose: I love the stream but I need more filters and bridges', *Gigaom,* 9 January, 2014 <https://gigaom.com/2014/01/09/drinking-from-the-twitter-firehose-i-love-the-stream-but-i-need-more-filters-and-bridges/>(18 September, 2017).

elections, the war on terror, Supreme Court nominations and the events of 11 September 2001.[69]

In comparison, all 21 billion public tweets from 2006-2010, had a total size of 21 terabytes. Its complexity is not particularly reated to its size, but by the fact the Twitter archive is so fractionated. To complicate the matter, the first 21 billion tweets came with 50 accompanying metadata fields. This first batch was delivered – by Gnip – a the beginning of 2012, but quickly followed by a second and even larger batch in December 2012. Consisting of 150 billion tweets worth 113 terabytes in data, the Library of Congress saw their Twitter archive expanding by 565% in less than a year. At that time the Library only held 167 terabytes of Web data, so the totaling 170 billion tweets and corresponding metadata were with a total of 133 terabytes almost on even ground. In the two years following the arrangement between the Library of Congress and Twitter, the platform experienced an enormous growth, that only stabilized around 2014.[70] The number of tweets by then had grown from 50 million at the time of arrangement in 2010 to a total of 500 million in 2014. As a result, Twitter had to change its architecture to keep their service running. Already in 2010 during the World Cup, the global conversation covering every aspect of the game, repeatedly took its toll and made Twitter unavailable for short periods of time. For an issue like this to be solved, Twitter had 200 engineers to address the problem.[71] This shows the possibility for Twitter to continually restructure their technological infrastructure and re-architecture how it processes, archives and displays its content and activity. If Twitter's structure had remained free of change, the first batch would have been sufficient to develop a cataloguing system, workable for all incoming tweets no matter the quantity. The Library of Congress simply has not got the workforce to keep up with a rapid developing platform like Twitter though. One of the contributing factors to illustrate this challenge is the growing number of metadata fields, that has expanded up to 160 fields metadata that can be associated within each tweet. So not only did the number of tweets grew, the tweets itself grew bigger too as users found ways to attach, photos, video's and live video streams.

[69] M. Zimmer, 'The Twitter Archive at the Library of Congress: Challenges for information practice and information policy', *First Monday, 20* (7), (2015) <http://firstmonday.org/article/view/5619/4653>(18 September, 2017).
[70] Zimmer, 'The Twitter Archive at the Library of Congress'.
[71] R. Krikorian, 'New Tweets per second record, and how!' *Twitter Blog,* 16 August 2013 <https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html>(18 September, 2017)

According to privacy and ethics scholar Michael Zimmer the challenge to archive the Twitter archive for the Library of Congress is a matter of both practice and policy. He argues that 'the practical challenges of receiving such a large volume of Twitter data, archiving it, and making it accessible and useful are sizeable, and undoubtedly the Library of Congress is putting forward great effort to resolve them quickly'.[72] It is the question however not if, but when this will happen. At the current state of being, it does not seem like much progress is being made. The latest report from the library is dated from 2013. In this report they admitted they had not find ways to properly deal with the challenges. At that time 'executing a single search of just the fixed 2006-2010 archive on the Library's system could take 24 hours'.[73] Also the library acknowledges that 'it is not uncommon for the Library to spend months or in some cases years sorting a large acquisition to inventory, organize and catalogue the information and materials so they are accessible by researchers.'[74] Time has shown that processing the Twitter archive is already a project that is taking years rather than months. According to the last news reports (2016) no engineers are permanently assigned to the project, resulting in a situation where 'staff simply dumps unprocessed tweets into a server – the digital equivalent of throwing a bunch of paperclipped manuscripts into a chest and giving it a good shake'.[75] In the article *Twitter as a First Draft of the Present* by Axel Bruns and Katrin Weller, published in the same period, this situation is confirmed. As they mention: 'to the best of our knowledge, the Library of Congress is obtaining the data via Twitter, Inc.'s subsidiary data reseller GNIP in a specific, textbased format, and is currently storing incoming tweets on various tapes (for tweets collected in a certain period of time) in a nonsearchable way'.[76] Zimmer states we should be wary of this process. The past has shown that when institutional organizations like the library lacked the resources to digitize all the books, only powerful companies like Google had the money and resources to make it happen.[77] The Library of Congress is heading in the same direction as Gnip, the company that

[72] Zimmer, 'The Twitter Archive at the Library of Congress'.
[73] Library of Congress, 'Update on the Twitter Archive At the Library of Congress', (2013) <https://www.loc.gov/static/managed-content/uploads/sites/6/2017/02/twitter_report_2013jan.pdf>(18 September, 2017).
[74] Library of Congress, 'Update on the Twitter Archive At the Library of Congress'.
[75] A. McGill, 'Can Twitter Fit Inside the Library of Congress?', *The Atlantean*, 4 August, 2016 <https://www.theatlantic.com/technology/archive/2016/08/can-twitter-fit-inside-the-library-of-congress/494339/>(18 September, 2017).
[76] A. Bruns and K. Weller, 'Twitter as a First Draft of the Present – and the Challenges of Preserving It for the Future', *Preceedings of ACM Web Science Conference,* (2016) <http://dx.doi.org/10.1145/2908131.2908174 >(7 November, 2017).
[77] Zimmer, 'The Twitter Archive at the Library of Congress'.

also delivers the tweets to the Library, currently offers the only complete and historical database for researchers.

Apart from the effort, money, time and expertise the Library is willing to use for the practical challenges, the policy challenges mostly address the issues of access and content restriction, privacy and user control. The content restrictions are defined in the agreement Twitter and the Library made in 2010 about the archive:

- It includes only public tweets;[78]
- The Library may display and otherwise make available public tweets only after a six-month delay;
- The Library will not provide a 'substantial portion' of the archive on its public Web site in a format that could easily be subject to bulk download;
- Access should only be provided to 'bona fide' researchers in accordance with 'the policies of the custodial division of the Library responsible for the administration and service of materials of this nature,' and only if the researcher signs a notification prohibiting commercial use and redistribution of 'all or a substantial part' of the archive.[79]

The rules regarding access restriction are clear, do not offer any real challenges as researchers are often enlisted to an university or research instruction, and are clearly set up to prevent any commercial use of the archive.

Content restriction offers are a more complicated challenge though. The Library has to honor the Code of Ethics of the American Library association that states its members are committed to 'intellectual freedom and the freedom of aces to information'.[80] Therefore all tweets should be freely accessible. Tweets however, can contain very personal and controversial information. Twitter itself regulates the platform and removes tweets that violates its policy, for example the hateful conduct policy that removes tweets that promotes violence against others based on race, national origin or gender.[81] This is done with the help of reports from users and therefore this method of content moderation is limited. According to the gift agreements the Library may dispose 'any part of the Collection' that 'is inappropriate for retention'. Thence, the Library is challenged to find the right balance between intellectual

[78] It is estimated that fewer that around 90% of all tweets are public. Zimmer, 'The Twitter Archive at the Library of Congress'.
[79] Library of Congress, 'Update on the Twitter Archive At the Library of Congress'.
[80] Amarican Library Association, 'Ethics' <http://www.ala.org/tools/ethics>(7 Novermber, 2017).
[81] Twitter Support, 'Our Policy' <https://support.twitter.com/articles/20175050>(18 September, 2017).

freedom and moderating tweets that might include hateful, copyright-protected, confidential or even illegal content. So far they have not published a policy for this issue.[82]

Challenges on the matter of privacy and user control appear because some rules that intent to offer privacy have a breach. The most efficient way to prevent a user's tweets from being published is by creating a private account. Yet, Twittering without an audience to tweet to is useless. The users within the audience of a private account have the opportunity to retweet though, resulting in a breach where private tweets can still end up in the archive. A similar issue where users experience a false sense of control is when they delete their tweets. It removes their tweets from the users feed and the Twitter search results. Deleted tweets or accounts only disappear from Twitter's database, but not from third parties and search engines.[83] They may still end up, however, at the Library's archive, because Tweets are delivered to the Library of Congress in static documents by Gnip, so what has been deleted after the delivery of a data batch will have reached the archive anyway. The Library responded to this issue by mentioning Twitter's terms of service users already agreed to and noting that the information is already public.[84] Yet, the issue remains whether the Library should not archive the 'breached' tweets in order avoid conflict with users or trust on the lawfulness of Twitter's terms of service. An update in these terms of service, stating tweets are archived by the Library of Congress, might be desirable in this case.

Apart from a privacy issue, deleted tweets also provide an issue considering the loss of valuable information according to Weller and Bruns. They may hold 'controversial statements by important public figures like politicians' and 'eye-witnesses of critical events', tweets that can be both withdrawn or deleted because offended other users. The two scholars also argue that, as the Library works with a text based format, another thing that the archive will lack is the 'Twitter's unique look and feel'. Chapter 1.1 has already explained how Twitter's structure has evolved. The inability to recapture this will make it hard to experience the early phases of the platform. Two other aspects that are at risk of being permanently lost are URLs and audiovisual information. While shortening URLs was helpful to create space in the 140 character limited tweets, adaptations of this shortening system or the impermanence of the linked sites both endanger the loss of the contextual information URLs provide. Audiovisual

[82] Zimmer, 'The Twitter Archive at the Library of Congress'.
[83] The Privacy Policy mentions 'Keep in mind that search engines and other third parties may still retain copies of your public information, like your user profile information and public Tweets, even after you have deleted the information from the Twitter Services or deactivated your account'. Twitter Support, 'Privacy Policy' < https://support.twitter.com/articles/20174657>(18 September, 2017).
[84] Zimmer, 'The Twitter Archive at the Library of Congress'.

content is not included at all, as both 'the free Twitter APIs' and the Firehose access through GNIP provide text-based information only.[85]

The Library of Congress could be the perfect counterpart to the commercialization and monopolization of social media data, however in its current situation this is merely an ideal far removed from reality. Twitter data is definitely not completely out of reach, but large quantities of its data are becoming harder to reach for scholars and might be lacking important contextual information. Therefore getting the data is the first challenge scholars have to overcome. This is however just the first step of a study. Once retrieved, the data needs to be structured, analyzed and read, before it can provide scholars with any answers. Therefore, the next chapter will elaborate on the various methods used by scholars, the results that are gained by these, and its potential for humanities, by discussing a number of studies.

Chapter 3: Turning Data into Knowledge

Peter Verhaar 30-11-17 17:03
Deleted: Page Break

In this chapter a number of studies using Twitter data will be analyzed to illustrate their potential and challenges for humanities. Two distinctions have been made to distinguish different kinds of methods and results used in Twitter studies. The first distinction is between the use of mainly metadata or textual data. This will provide an overview of the variety of knowledge that can be extracted from Twitter data. The second distinction, is made to emphasize studies that *apply* concepts to analyze, structure and interpret the used Twitter data, compared to studies merely process data into visualizations, without further explaining or interpreting their results.

Although the goal of this thesis is to analyze the potential and challenges of Twitter data for humanities, the majority of the analyzed studies do not directly fit within the discipline of the humanities. They do fit within disciplines such political science and psychology, studies that are closely related to the humanities, which also study aspects of human society and culture. The studies are chosen because they either create results humanists can *exract* knowledge from such as maps, social networks and collective identities, or use methods humanists can *apply* to twitter data to study human society and culture.

3.1: Studies based on Metadata: Mapmaking and Communicative Networks

Collecting the data is almost always the first step for a data-intensive study, but the type of data collected is often in consideration with the methods scholars want to apply to it. Tweets

[85] A. Bruns and K. Weller, 'Twitter as a First Draft of the Present'.

appear to be mainly textual, but as mentioned they can contain over 160 fields of metadata. Therefore studies can be executed without the use of the content of tweets, or they may combine textual data with meta-data. In this first paragraph, a couple of studies will be explored that rely heavily on metadata, to see which methods are used and how these provide results that can be turned into knowledge for humanities scholars.

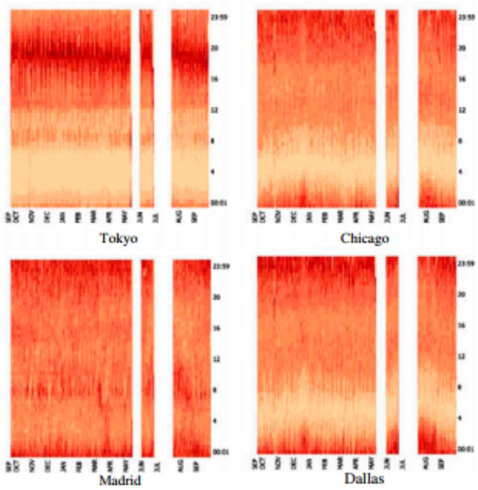


Figure 3: Four heat maps from the study by Adnan, Leak and Longley. Source: 'A Geocomputational analysis of Twitter activity around different world cities' (2014).

Twitter itself is not unwilling to get their hands dirty on its own data. Analyses of the Twitter data set offers the company the opportunity to explore the value and potential of what it rightfully owns, and to exhibit visually impressing datasets on its own platform. On *Twitter Interactive* it portrays visualizations of heavily debated aspects of popular culture like the Game of Thrones TV-series, the Oscar nominations and FIFA World Cup, political matters like the development of the U.S. presidential elections and geographical visualizations of political, musical and sport preferences around the world.[86] Here, various examples can be found of how a combination of merely metadata can lead to cultural knowledge. One of them is the visualization of Twitter activity patterns for four different cities; New York City, Tokyo, Istanbul and Sao Paulo. By combining the number of tweets, time of posting and the location, the activity patterns of these cities can be visualized with a heat map, showing when as city is active or asleep. The varying patterns detected in these visualizations indicate cultural differences. For example Tokyo does not show any seasonal changes in its activity pattern, whereas all other cities show a spike in late night activity in July and/or August.[87] This visualization can help those who study Japanese culture, for example to describe and debate the presence of a high and consistent work pressure in Japan. Twitter's own research analysts do not burn their fingers on any attempt to explain or interpret their findings, though.

[86] For more examples, see: Twitter, 'Twitter interactive' <https://interactive.twitter.com/>(17 October, 2017).
[87] J. Linn, ,'Studying rapidly evolving user interests', *Twitter Blog,* 4 June 2012 <https://blog.twitter.com/engineering/en_us/a/2012/studying-rapidly-evolving-user-interests.html>(17 October, 2017).

This study however, has been replicated and expanded by the geographical scholars Adnan, Leak and Longley, who visualized the activity patterns of the top 15 Twitter cities. They make a minor attempt to explain local differences, but their main focus is handling the data instead of turning into knowledge. Compared to patterns from Twitter's analysts, they have one shortcoming, as they are missing data between the months May and June and July and August. This illustrates the complexity for researchers without a business relation to Twitter to retrieve a complete dataset, or fill in the missing data without any extra cost.[88] For humanities scholars, nonetheless, it has the potential to lead to indications of cultural differences and to explore factors that cause the difference like religious celebrations, working culture and night life. They can fill the gap between data visualization and knowledge, although the incomplete dataset might fail to visualize important cultural events. It is preferable to have complete datasets of multiple consecutive years, to determine the recurrence of cultural events. Therefore, the potential is limited, when the tools to create a complete dataset are unavailable.

Geographic information like demographics have always helped humanities scholars like historians to explain political, cultural and religious differences. Yet, these boundaries are often defined by government agencies for administrative economic and political purposes. Twitter's geographical data offers rich opportunities for geographers and the ability to 'delineate non-administrative anthropographic urban boundaries by constructing a mobility network of Twitter user spatial interactions', which resulted in studies in which Twitter data was used to redistribute the mobility networks, urban boundaries and socio-economic relationships in Great Britain and London.[89] This demonstrates how Twitter data attributes to innovative methods for mapmaking. The detailed, crowd generated, geographic information assembled by the platform also enables scholars to describe and redefine location, as advancing mobile social media technology changes the experience of location-based information for its users. Twitter is a locative platform, as it is interested in where you are and offers the opportunity to show local trends. For its users this means they receive information 'wherever they are', which is 'increasingly about where they are'.[90] Informative local information, based on personal interest, will lead to more engagement of Twitter's users and

---

[88] M., Adnan, A., Leak & P. Longley, 'A Geocomputational analysis of Twitter activity around different world cities', *Geo-spatial Information Science,* (2014), p. 1-8.
[89] J. Yin, A., Soliman, D., Yin, S., Wang, 'Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data.' *International Journal of Geographical Information Science, 31*(7), (2017), pp 1219-1313.
[90] R. Wilken, 'Twitter and Geographical Location' in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014), p. 156.

directed advertisement, which is therefore desirable for Twitter's business. This growing locative interest of not only Twitter, but also other mobile media applications, combined with the popularization of location-aware mobile technologies has forced scholars to rethink the traditional conceptualization of location. 'Locations are still defined by fixed geographical coordinates, but they now acquire dynamic meaning as a consequence of the constantly changing location-based information that is attached to them'.[91]

Studying geographic Twitter data serves multiple purposes each with their own approach. Firstly, it provides informational data that allows geographers to create innovative demo- and ethnographics, used to create maps and charts. Studies like the one applied to Great Britain and London, often combine various fields of metadata to increase the number of geo-located tweets, which is generally about 1-2%. With the help of language detection, information from Twitter biographies or analyzing places mentioned in the tweets, the amount of geo-located tweets can be extended. Hereby the textual content is being used to create the metadata for researchers with the help of algorithms. A combination of such algorithms are used in the study *Mapping the global Twitter heartbeat: The geography of Twitter*, a study that is highly focused on creating a methodology suitable for the creation of an increasing and more accurate amount of data.[92] These methodological advances, fueled by technology, make way for a second approach, which is reflecting on how it is reshaping our media use and experience.[93]

Apart from mapping geographical metadata fields, mapping social communities, such as political, scholarly or Twitter influencers networks is a popular topic for studies working with metadata. These communicative networks can be visualized by analyzing the followers-followees network from multiple users to see how it is concentrated. To illustrate this method, the Twittering community of digital humanists visualized their own 'community of practice' on Twitter.[94] To identify users that are part of the digital humanities in the first place, they used various approaches. The most visible users (established professors and researchers in the field) served as a top-down starting point. Their followers were reviewed and selected

---

[91] R. Wilken, 'Twitter and Geographical Location'.
[92] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan & E. Shook, Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday, 18*(5), (2013). <http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654>(18 October, 2017).
[93] Studies such as *A Mediatized World*, elaborately review how our daily interaction with new media technologies changes our lives. A. Hepp& F. Krotz, *Mediatized worlds : Culture and society in a media age.* (Basingstoke, Hampshire: Palgrave Macmillan, 2014).
[94] M. ‚Grandjean, 'A social network analysis of Twitter: Mapping the digital humanities community', *Cogent Arts & Humanities, 3*(1), Cogent Arts & Humanities, Dec 2016, Vol.3(1) (2016).

according to their Twitter biography. In addition those who contributed to discussions around digital humanities related conference hashtags were added and more followers were gathered around key term related searches. The visualized community of practice, shows a clustered cloud at first, but with the addition of the main language tweeted in by the users, four different communities (English, French, German and Spanish) can be spotted. The language sub groups show that the French are somewhat isolated, hinting at the fact that community members might find it harder to interact with or read tweets form the French members. Apart from language the study analyzes the aggregation around influential members of the community.[95] Analyzing a scholarly communication and connections is a traditional theme for humanities scholars. Mapping networks has also been applied to the intellectuals from the Enlightenment in the project mapping the *Republic of Letters* from the Standford University.[96] Likewise, Twitter can now be used to show the central persons in a scholarly network and its development in real-time. Also, it provides the opportunity to ask a whole new set of questions such as the following: What is the influence of the amount of followers on article references? What are recurring topics in conference hashtags? Are French academics more isolated in general? These are all matters the data hints at, and can be studied with more Twitter data.

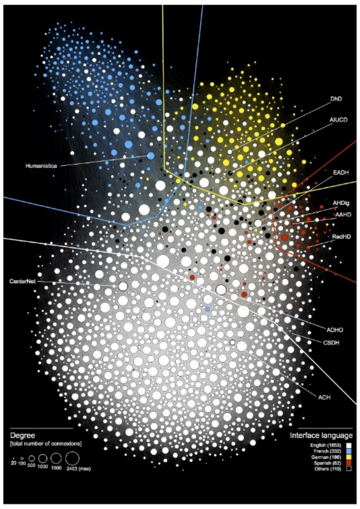


Figure 4: The cluster of the Digital Humanities network (white = *English, blue = French, yellow= German, red= Spanish)* Source: 'A social network analysis of Twitter' (2016).

Working with networked communities can also give insights into the online strategies of politicians. In the study *The Gift of the Gab* the tweeting behavior of all Dutch members of parliament between the period of 1 February and 31 August is analyzed in two different manners. Firstly, by their replying behavior and secondly by their retweeting behavior. With the help of the visualization program Gephi, members are given colors depending on the party they belong to. With interconnected nodes all the replies and retweets are visualized in a web. The reply network has a very round shape, which indicates the parliament members frequently communicate with each other on Twitter regardless of their party affiliation. The

[95] M. ,Grandjean, 'A social network analysis of Twitter'.
[96] Various representations of scholarly communication networks can be found on Stanford University, 'Mapping the Republic of Letters' <http://republicofletters.stanford.edu/>(18 October, 2017).

retweet network shows clear and demarcated clusters though. Apart from the clustered parties, the political color and the ruling government is also visible, as the right wing and governing parties form clusters on the right side, while the left-wing parties form clusters on the opposing side. It visualizes how replying and retweeting serve different political tactics. Replying is a form of communication and debate in which politicians can express their own opinion by debate even when their views are opposed the view expressed in the tweet they reply to. Retweeting is more often a form of endorsement and is therefore mainly applied to the tweet of party colleagues, as they share the same beliefs and values. The study explains these different practices, by explaining them as part of a human tendency, the concept of homophily:

> 'Similarity breeds connection. This principle —the homophily principle— structures network ties of every type... The result is that people's personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics. Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience. (McPherson, Smith-Lovin, & Cook, 2001, p. 415)'[97]

By adding this concept and applying this to the visualized network, this study distinguishes itself from the data studies that have been analyzed in this chapter so far, which have the tendency to focus on the methodological structure and data of a study. In combining concepts and data, humanities scholars can prove themselves valuable, as they have carefully put together these concept by studying human behavior, sociodemographics and intrapersonal characteristics. Network ties like political parties that rely on this principle, are visible in real life as they express themselves in organizations. Yet, visualizing these networks with the help of Twitter helps to verify a concept like homiphily, which provides insight in human tendencies. Homophily is commonly described with the phrase 'birds of a feather flock together'. With the help of social media networks and powerful visualizations, scholars can now see flocks of people as clear as flock of birds in the sky. If Twitter will be successfully archived by either the Library of Congress, or a different (commercial) institution, Twitter will provide future historians with the potential to recreate real time events and connections, network analysis can be used to study how communities of people online form and dissolve and study the cause and effect.

[97] Quote from: McPherson et. al, 'Birds of : Homophily in social networks' in J. Paßmann, T. Boeschoten and M.T. Schäfer, 'The Gift of the Gab: Retweet Cartels and Gift Economies on Twitter' in in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014), p. 336.

3.2 Studies based on the Textual Content of Tweets

Studies focusing on followers-followees networks can display communities in the form of networked individuals. Studying the content of tweets, especially under certain hashtags, can be very effective to identify communities that share the same values and belief. As Twitter is often said to serve as 'the world's largest villages square' for debates, communities with opposing views can often be identified through the fact that they use distinctive hashtags. This way, Twitter users are able to participate in public discourse and shape it.[98] The 2016 Presidential Election in the US has provided scholars with a rich data pool to study the expression of conflicting thoughts under different types of hashtags. With nearly 70 million people monthly active on Twitter[99], and a system where two candidates of opposing parties strove for the presidential seat, an enormous amount of conflicting political expressions can be found on Twitter. Political supporters of respectively Hillary Clinton and Donald Trump express themselves under opposing hashtags to show their political preference. Apart from showing their support with hashtags like #ImWithHer and #TeamTrump, users also used anti-hashtags like #NeverHillary and #NeverTrump. This provides scholars with the opportunity to reconstruct the ideas and values for both groups, by analyzing all their expressed opinions in tweets.

However, with the help of supporters writing tweets containing these hashtags, it is not only possible to analyze the political views of Twitter users, but also to study their personality type. In the study *Personality and Politics* Juola and Vinsinck used the hashtags to define four groups: pro-Clinton, pro-Trump, anti-Clinton and anti-Trump. Around 600 subjects from each group were randomly selected to analyze. The study used the individual profiles of these subjects to carry out the Myers-Brigss Type Inventory (MBTI), a well-known personality test. This was done with the help of EthosIO, an API that analyzes people who write text, where text is this case was the full Twitter feed of the selected persons. Although the overall statistics did not match US demographics as some personality types were strongly overrepresented, they did find significant personality differences between 'Democrats' (pro-Clinton and anti-Trump groups) and 'Republicans' (pro-Trump, anti-Clinton). Despite the finding that the personal feeds were not sufficient to create accurate personality tests, this study forms a good illustration of how Twitter data can contribute to new insights considering the motivations of voters. The personally differing psychological preferences of how people

[98] P. Juola and S. Vinsick, 'Personality and Politics: Myers-Briggs Personality Types on Twitter in the US 2016 Presidential Election', *Digital Humanities Congress 2017 Abstracts*, (2017).
[99] Twitter Files, 'Selected Company Metrics and Financials'.

perceive and participate in the world around them, can be a valuable attribution to the more traditionally social, political and economic reasons that motivated voters.[100]

The method applied in *Personality and Politics* can be considered as a survey executed without the users' consent. As Twitter users express themselves online, and hashtags serve as a topical filter, the platform provides scholars with the opportunity to carry out surveys regarding specific topics. In *#ww1. The Great War on Twitter*, contemporary historian Fréderic Clavert uses Twitter to research the collective memory of the Great War in England and France. The data was collected with streaming API. The study collected Tweets from the 1st of April 2014 and onwards. As the Great War started and the end of July in 1914, it was by then 100 years ago, which lead to increased attention for the Great War in Europe. The study is a good example of how the streaming API is adequately put to use, to analyze a planned and enhanced commemoration of an event that shaped the world. It lead to a total collection of around 1,5 million tweets that contained a hashtag that was related to the Great war, written by 350.000 different Twitter users. The majority of the tweets were in English (90%) with French being the majority in the non-English tweets. With the help of textual analysis two major differences were found in the way the English and the French commemorated the great War. The French mainly focused on the soldiers, the *poilu* (a name only given to French WW1 soldiers), and the end of the War, while the English also concentrated on battles and how Great Britain entered the war. With this study Clavert basically had the opportunity to asks thousands of people 'How do you remember the Great War' without really asking them, to construct and compare the collective memory of two countries. Beside this textual approach that used tweets, the study also carried out social network analysis and network visualizations with the help of metadata.[101] By doing so it did not only investigate how the Great war is remembered, but also how Memorial and Heritage Institutions and both professional and amateur historians a play the role in inciting the memorial culture.[102]

3.3 Applying Concepts to Textual Data

These two studies show how textual Twitter data can be used to create new knowledge directly from data, in the form of voters' personalities or collective memory. In this thesis the

[100] P. Juola and S. Vinsick, 'Personality and Politics'.
[101] This is only briefly mentioned in the abstract, the focus throughout abstract is on *how* the great war is remembered in tweets. Yet, it does show how one corpus can combine the results from both textual- as metadata. F. Clavert,, '#ww1. The Great War on Twitter', *Digital Humanities Congress 2017 Abstracts*, (2017).
[102] F. Clavert,, '#ww1. The Great War on Twitter'.

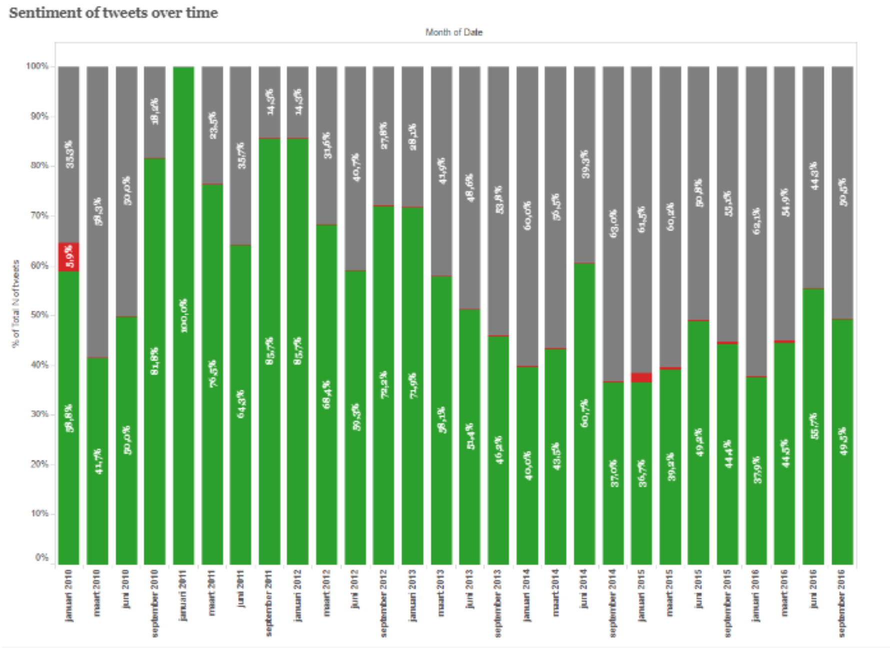


Figure 5: The sentiment of Wilders' tweets over time (Grey = neutral; Green = supporting new right; Red = opposing new right). Source: 'Analysing tweets of Trump and Wilders'.

evaluated studies that worked mainly with metadata showed how the concept of homophily is used to *interpret* the network visualizations. In *Personality and Politics*, the concept of the MBTI is *applied* to study the relation between political preference and personality. In this paragraph, studies will be evaluated that *apply* concepts to structure and analyze textual data. This is a common approach for studies working with textual data, for example in a study by the Digital Methods Initiative, which analyzes the tweets of Trump and Wilders. In this study an attempt is being made to track the concept of populism in the tweets of Wilders and Trump. The concept populism in this study is perceived as 'a political discourse that imagines a struggle between good and virtuous people: and a nefarious establishment', whereas a populist leader is defined as 'a charismatic leader who uses this kind of thinking to mobilize large numbers of people to gain and hold power. Populists can be either on the left or on the right; the outlook combines with a variety of other ideologies or issues'.[103] The study identifies Alt-Right as a populist right wing movement in the U.S. and New Right as its European counterpart. Although both movements address different issues, as they act on

[103] M. Blonk, V. Buss, M. Scherf, J. Voorn & P. Vliegenthart, ,'Analysing tweets of Trump and Wilders', *Digital Methods Initiative, Winter School 2017*, (2017). <https://wiki.digitalmethods.net/Dmi/WinterSchool2017TrumpWildersTweetAnalysis>(18 October, 2017).

different continents, both groups have several matching ideologies. They oppose globalization, favor isolationist foreign policies and oppose immigration and Islamification. In order to mobilize these groups and win their votes, Trump and Wilders can reproduce the ideologies of respectively the Alt-Right or the New-Right movement in their tweets.[104]

As both leaders are very active on Twitter, the data set contained 3,644 tweets from Trump and 6,470 from Wilders. As a first step, the study classified the tweets. For Trump this meant his tweets were analyzed and discussed, to investigate whether he mentioned themes from the Alt-Right movement like 'Political correctness', 'Mainstream Media', 'Gun Control' and 'Black Lives Matter'. For Wilders tweets the same approach was used, but with categories in line with the New Right movement like 'Establishment', 'EU', 'Freedom of Speech' and 'Immigration'. This classification system was derived from other studies and news articles that described both movements and their characteristics. By doing so, the study applies these concepts to structure the textual data. The categories are then used to research to what extent they are mentioned in the tweets. If a tweet holds information about something irrelevant considering these categories, it is considered to be neutral. Yet, a tweet is also neutral when it concerns one of the categorized issues but neither supports or opposes it.[105]

The first and most general results this approach provided for the study, are the percentages of tweets considered as neutral for Trump (77,1%) and Wilders (51,9%), or opposed the views of Alt-Right (Trump, 2,63%) or New-Right (Wilders 0,28%). This can be plotted against time in months to see in which period both leaders exported a more 'populist' discourse on twitter. This plot only displays the three variables neutral, supporting and opposing. The supporting tweets however, can be divided into the classifications made for both the Alt-Right and New-Right movement. This way the topical coverage of each political leader can be plotted against time. Combined these graphs are able to indicate the major turning points in both political careers. These major changes in sentiment that are displayed by analyzing their Twitter feeds, are further explored and explained by examining the events they are related to. One example from the study is the drop in the percentage of neutral tweets, when it became clear that Trump most likely became the Republican nominee. For Wilders the study displayed a different trend, as he tends to send out less tweets that are identified with the New-Right approaching the elections.[106] With the visualization of the

[104] Blonk et. al,'Analysing tweets of Trump and Wilders'.
[105] This visualization is made in an interactive graph and can be viewed at: Blonk et. al,'Analysing tweets of Trump and Wilders'.
[106] Blonk et. al,'Analysing tweets of Trump and Wilders'.

sentiments tweeted by both political leaders, the study creates insight in political tactics and considerations made by both regarding their Twitter behavior. It shows how much they address the topics of right-wing populist movements over time, but not necessarily if they identify with these groups, because they never directly admit their solidarity to these groups.

Detecting and evaluating a populist discourse is not new for humanities scholars, but the method used is. Basically, it distinguishes itself by the method of distant reading compared to the traditional method of close reading. Careful analysis of traditional media like newspapers and television could have led to the same assumption of Trump and Wilder as populist politicians. The Twitter data has the potential to measure the concept of populist discourse by the both of them, to compare these and to spot changes over time. Apart from the methods, the percentages and graphs created by the study can help humanities scholars to further explore political tactics and the concept of populism.

Political topics are the most popular subjects of conversation on Twitter, especially in the United States. They generate more comments than non-political posts.[107] Despite the vast number of political tweets, close reading is still applied in Twitter research. In *#refugeesnotwelcome: Anti-refugee discourse on Twitter* Ramona Kreis analyzes 100 tweets manually. The tweets included the hashtags #refugeesnotwelcome and were collected in a two-day timeframe in September, two days after Germany had closed its Southern border. For the selection of qualitative data in the corpus, guidelines were followed. The sample was collected from 'thematically organized streams of discourse'.[108] The goal of the study was to check how immigrants and refugees are represented in tweets that include #refugeesnotwelcome and to study which discourse features are used in these tweets. In this study the discourse features are elaborately explained, to illustrate what Kreis is studying in the tweets. One of the applied concepts is critical discourse, which analyzes how 'discourse (re)produces social *domination,* that is the *power abuse* of one group over others'. With regard to this study, analyzing critical discourse might reveal how nationalist group might create an 'in-group' (European/Christian/homogenous) versus an outgroup (refugees/immigrant/foreigners), where expression like 'our' (land/place/country) are used to

[107] R. Kreis, '#refugeesnotwelcome: Anti-refugee discourse on Twitter', *Discourse & Communication,* Vol. 11 (5), 2017, pp. 498-514, p. 501.
[108] J. Androutsopoulos Online data collection, (2013). In: R. Kreis, '#refugeesnotwelcome: Anti-refugee discourse on Twitter'.

set the boundary of the in-group This way Kreis studied the rhetoric of those who contributed to this hashtag, and enhanced their positive self- and negative other representation.[109]

Apart from critical discourse aspects, the study also applied the multimodal discourse concept, which focuses on how language is used. This concept analyzes how we give meaning to language. This varies whether communication is expressed in written form or in spoken language. When speaking, intonation, voice quality and rhythm are aspects that influence the perceived meaning. On Twitter however, users are limited to typographic elements, like using capital letters, repetition and emoticons, but also anything they attach to their message like blogs and videos. One of these multimodal discourse aspect is already present in each tweet, namely the #refugeesnotwelcome, among which users can 'bond around particular values'.[110] In addition, hashtags like #GOHOME or #remigration can be added to adhere further to this point of view.

The study analyzed the tweets and spotted a wide variety of aspects that fall under the concepts of both critical discourse and multimodal discourse. Therefore it provides insight in how rhetoric on Twitter is used to create contrast between the in- and out-group. Kreis argues that 'the data reflects the growing sentiments against immigration and refugee policies and practices in some part of European societies, which seems to point out to the rise of an ideology of White dominance and superiority as well as nationalism and right-wing populism in Europe'.[111] This conclusion might be somewhat premature though, especially considering the growth she perceived. Unlike the study that analyzed populism in Wilders and Trump tweets over a period of years, the data from this study was a reflection of only two days. To study a growth in sentiment a combination of methods from both studies might be more suitable. The concepts of Kreis considering modes of discourse can hereby serve as a classification system for sentiment analysis, to detect a growth in anti-refugee sentiment. Therefore historical analysis of refugee crisis related hashtags is necessary though. As explained in chapter 2, it can be difficult and costly to retrieve such data. This illustrates both the possibilities and the challenges humanities scholars experience regarding Twitter data studies. The classification method and sentiment analysis can be used for Trump and Wilders'

[109] : R. Kreis, '#refugeesnotwelcome: Anti-refugee discourse on Twitter', p. 502.
[110] M. Zappavigna *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*, (2012). In: R. Kreis, '#refugeesnotwelcome: Anti-refugee discourse on Twitter'.
[111] R. Wodak and S. Boukala 'European identities and the revival of nationalism in the European Union: A discourse-historical approach', (2015b) . In: R. Kreis, '#refugeesnotwelcome: Anti-refugee discourse on Twitter', p. 511.

accounts because the REST API allows researchers to download a user's feed, but the REST API cannot freely provide a historical set of a large group of users regarding the #refugeesnotwelcome.

3.4 Processing data versus interpreting data

Although only a handful of studies were evaluated, different approaches of Twitter data and its potential can be distinguished. With the help of metadata maps or social networks can be created. Hereby, the data is processed into sources of visual information, useful for humanists to ask or answer different questions about human culture and society. Studies such as these can be used to *extract* knowledge from. Within these studies, however, the focus in on processing this data into visualizations, rather than interpreting it to discuss aspects of human society and culture. This creates the potential for humanists to further interpret and explain these visualizations.

The studies *#ww1. The Great War on Twitter* and *Personalities and Politics* demonstrated Twitter's potential of taking surveys with the help of textual data. Especially *#ww1. The Great War on Twitter* demonstrated how the Streaming API was sufficiently used to analyze contemporary historic narratives of the Great War.

The studies that *applied* concepts, mainly focused on the political views of Twitter users. By applying concepts, these studies work with a broader theoretical framework, suitable to interpret and structure the analyzed data, allowing them to define the discourse of different sub groups of users. With the help of a broad theoretical framework, gathered by analyzing various Twitter studies, the next chapter will focus on one of the biggest social movements on Twitter: the Black Lives Matter movement, which focuses on racial issues in the United states.

Chapter 4: A Case study; #BlackLivesMatter, an echo chamber or conflict space?

Chapter 3 demonstrated that Twitter data can provide insight in different political views and movements. It is a challenge though, to find the right corpus, preferably one that analyzes how opinions evolve over a long time. Therefore, this case study will attempt to combine multiple studies, to compare and evaluate whether echo chambers are persistent on Twitter. Hereby the results from other Twitter studies regarding are *extracted* to ask new questions. Concept such as the echo chamber will be carefully evaluated to see if they *apply* to corpus gathered for the case study. Throughout the case study, the focus on the potential and challenges of working with Twitter data will remain a central topic of discussion.

4.1 Studying the most influential hashtags: #Ferguson and #BlackLivesMatter

The study *Ferguson and the death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the evolution of collective identities* analyzes the concept of collective identities ,and how opinions evolve within it during four meaningful periods within a year. The study collected 31.65 million tweets about Ferguson where Michael brown, an 18 year old black man, was fatally shot by a white police officer on 9 August 2014. The study used both the streaming and REST API to collect tweets that mentioned the term ‘Ferguson’. As it started collecting data only five days after the shooting incident, the study had to use the REST API to capture the tweets in the period directly after the shooting. This led to a situation where the tweets from 9 and 10 August, days critical to evaluate the first form of diffusion and reception of the news, initially could not be collected. This exemplifies the difficulty to capture the tweets posted at the origin of an event. Even when the REST API is operated within the seven day time frame, a situation might arise in which tweets are not retrieved from Twitter’s server, without any explanation why, because the algorithms’ actual functioning is known only to Twitter. Eventually the study was only able to retrieve these tweets from these two days via Twitter’s search interface, that allows you to search for any tweet that has ever been posted with the help of queries. By clicking on a tweet, its ID number is displayed in the browser bar. Up to 100 tweets per request can be retrieved by entering this ID as a comma separated value through the GET satuses/lookup API.[112] Hereby the study demonstrates that there is an alternative way to retrieve, or ‘hydrate’ as this procedure is called, tweets outside of the REST API’s seven day window. Yet, it requires at least the computational skill to

[112]Twitter Developer Documentation, ‘GET Statuses/Lookup’ <https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup>(25 October, 2017).

program a bot to collect al IDs, or a very labor-intensive manual collection, as they were able to collect 32,056 ID's in total.[113]

Table 1. Number of tweets per period.

Death of Mike Brown (18 days)	13,238,863
Non-indictment of Officer Wilson (28 days)	15,080,082
DoJ Report on Ferguson (8 days)	2,033,898
One Year Aftermath (13 days)	1,304,702
Total	31,657,545

Figure 6: Periods of Data Collection. Source: *Ferguson and the death of Micheal Brown on Twitter* (2017).

After the study evaded the REST API's limits, the vast number of tweets over a period of a year provided the study with enough possibilities to study 'what collective identities (in the form of hashtags) emerge and survive over time as they relate to Ferguson on Twitter'. The second research question for the data set was to locate 'the themes that are linked to surviving and collective identities on Twitter'.[114] Although the tweets were collected over the course of a year, they were only collected during four periods were a volume of activity (which was undefined in the study) was perceived. These periods were centralized around key developments in the Ferguson case, as is shown in table 1 from the study. The aftermath of Mike Brown's death established #BlackLivesMatter as an online social movement. Before the incident it was only slowly gaining prominence on Twitter, after the phrase was introduced by Alicia Garza in a Facebook post called 'a love letter to black people'.[115] The post was a reaction to a similar incident; the death of the seventeen-year-old Trayvon Martin who got fatally shot by the neighborhood watch volunteer George Zimmerman on July 13, 2013. Despite the fact that #Ferguson and #BlackLivesMatter are good for the first and third most used hashtags ever, Travon Martin's death only led to a total 5,106 tweets with #BlackLivesMatter in the second half of 2013.[116] Although the actual shooting of Mike Brown resulted into a quite similar volume of tweets mentioning Ferguson, the non-indictment of officer Wilson three months later marked the real rise of the #BlackLivesMatter movement. In the weeks after Brown's death, #BlackLivesMatter was only the 57th most popular hashtag in the lively Ferguson discussion, but after the non-

[113] R. Ray, M. Brown, N. Fraistat & E. Summers, 'Ferguson and the death of Michael Brown on Twitter: #BlackLicesMatter, #TCOT, and the evolution of collective identities', *Ethnic and Racial* Studies, 40: 11 (2017), pp. 1797-1813, p.1801.
[114] Ibidem, p. 1798.
[115]Garza ended her letter with 'black people. I love you. I love us. Our lives matter'. Her friend Patrisse Cullors amended these last three words into the hashtags #BlackLivesMatters. J. Cobb, 'The Matter of Black Lives', *The New Yorker,* 14 March 2016 <https://www.newyorker.com/magazine/2016/03/14/where-is-black-lives-matter-headed>(25 October, 2017).
[116]M. Anderson and P. Hitlin, 'Social Media Conversations About Race', *Pew Research Center*, August 2016 <http://www.pewinternet.org/2016/08/15/the-hashtag-blacklivesmatter-emerges-social-activism-on-twitter/>(25 October, 2017).

indictment of officer Wilson it had established itself as the third most popular hashtag, with 189.210 tweets on the day after the non-indictment at its peak.[117] In the two remaining periods it kept is top position, as is shown in the table.

These numbers demonstrate that the #BlackLivesMatter indeed emerged in the year following Michael Brown's death, but not necessarily why it emerged at that moment. The study *Ferguson and the Death of Michael Brown* used the data primarily to reconstruct the growth of the #BlackLivesMatter movement. By analyzing every randomly selected fiftieth tweet form the hashtag, they categorized the tweets regarding their content. This resulted into seven different categories, which are displayed in table 3 from the study. Twitter users attributing to #BlackLivesMatter firmly formed a collective identity, mainly over the topic of 'blacks being killed with impunity and whites not' and 'displays of solidarity and activism'. The study's aim was to investigate whether there was a counter narrative, forming an opposing collective identity. Both mainstream media and the Pew Research Center portray #AllLivesMatter as a counter narrative, as there is a broad collection of articles that compare both hashtags and an extensive comparison of the tweets from both hashtags by the Pew Research Center.[118] Yet, the study found that #TCOT (Top Conservatives on Twitter) turned out to be the counter narrative in the tweets collected around the term Ferguson. The #TCOT is popular simultaneously, but most importantly it seems to operate as a direct reaction to #BlackLivesMatter, by creating 'validating justifiable homicides' as its most popular theme. The study argues that 'these hashtags serve as polarizing collective identities about race and policing in America'.[119]

Table 3. #BlackLivesMatter themes.

	Percentage	Number
Blacks Killed with Impunity and Whites Not	43.24	224
Displays of Solidarity and Activism	34.36	178
Historical References to Discrimination	6.18	32
Demands for Policy Changes	5.60	29
Response to Race Card Claim	4.63	24
Media Double Standard	3.86	20
Humanizing Police Brutality victims	2.12	11
Total	100	518

Table 4. #TCOT themes.

#TCOT themes	Percentage	Number
Validating Justifiable Homicides	26.25	110
White Victims of Black Criminality	21.72	91
#BLM as Radical Terrorists	14.80	62
Black Problems for Black People	14.32	60
Media Double Standard	11.69	49
Tokenizing Examples of Blackness	8.35	35
Humanizing Police Officers	2.86	12
Total	100	419

Figure 7: Narratives from both hashtags extracted from the data. Source: *Ferguson and the death of Micheal Brown on Twitter* (2017).

---

[117] Ibidem.
[118] A collection of these articles can be found here: Huffpost, 'All Lives Matter' <https://www.huffingtonpost.com/topic/all-lives-matter>(25 October, 2017) and the tweets are analyzed in Pew Research Center, 'Social Media Conversations About Race'.
[119] R. Ray et. al. 'Ferguson and the death of Michael Brown on Twitter', p.1807.

Hereby the study contributes to the debate on whether social media, and therefore Twitter, deepens the political polarization in the United States. Recurring concepts in this debate are the filter bubble and the echo chamber. The filter bubble is created by algorithmic functions that calculate your interests and preferences and adapt your online experience accordingly. Thereby it narrows your online feed, creating echo chambers in which users already existing opinion's and beliefs are repeated, rather than contested or exposed to any opposition. [120] These processes enforce the human tendency to 'search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses', known as the confirmation bias.[121] Indeed, the study detects two opposing collective identities, but is hard to link the existence of these opposing identities directly to Twitter. Polarization might be expressed and perceived online, but caused by other, traditional media platforms. Research shows examples of how traditional media also adds to polarization by publishing one sided left-wing views, ignoring or setting aside any form of critic from the right-wing.[122] For general internet use, research shows that in the US polarization has largely increased among people who are less likely to use the internet, making it hard to point at the internet, and thereby social media, as a polarizing force.[123]

---

[120] E. Pariser, *The Filter Bubble : what the internet is hiding from you* (London, 2011).
[121] Wikipedia, 'Confirmation Bias' <https://en.wikipedia.org/wiki/Confirmation_bias>(7 November, 2017).
[122] S. Lindhout, 'Onderzoek: Duitse kranten namen Merkels 'Wilkommenskultur' vrijwel kritiekloos over', *De Volkskrant,* 24 July 2017 <https://www.volkskrant.nl/buitenland/onderzoek-duitse-kranten-namen-merkels-wilkommenskultur-vrijwel-kritiekloos-over~a4507954/>(7 Novermber, 2017).
[123] L. Boxell, , M. Gentzkow and J. Shapiro, 'Is the Internet Causing Political Polarization? Evidence from Demographics', *NBER Working Paper Series*, 2017 <https://www.brown.edu/Research/Shapiro/pdfs/age-polars.pdf >(7 November, 2017).

Longitudinal research of Twitter users, however, indicates that on this platform in the US polarization increased by 10-20 percent between 2009 and 2016. These percentages were derived from analyzing personal accounts and whether they became less likely to follow, retweet or use hashtags from both sides of the political spectrum (left wing and right wing).[124] These results confirm an increasing filter bubble and echo chamber. Concerning the study regarding #Ferguson however, it is more complicated to apply the concept of an echo chamber. The opposing #BlackLivesMatter and #TCOT were collected around the neutral term Ferguson. Besides, the term takes form in multiple neutral popular hashtags during all four periods, for example; #fergusondecision (period 2, rank 4), #fergusonreport (period 3, rank 3) and #worldwatchesferguson (period 4, rank 5).[125] With the exception of period 3, #mikebrown is the most popular hashtag in the range of the term Ferguson. Ferguson, the hashtags including Ferguson and #mikebrown are all neutral terms, they are used to refer to the shooting without making a political statement. Within these terms however, the opposing #BlackLivesMatter and #TCOT operate and clash. The echo chamber in this case only appears in users personal feeds, at a meso level, where they follow like-minded people confirming to the filter bubble concept. If Twitter users follow the continuous stream of tweets under the neutral terms at a macro level, they are confronted with conflicting opinions.

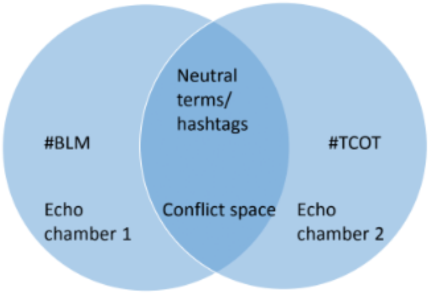


Figure 8: A visualization of clashing identities. Twitter users are in their safe space following their meso network or preferred hashtags, but are confronted with conlficting opinions when following the neutral terms and hashtags.

With the Ferguson study as prime example, Twitter data has proven itself useful to study collective identities, opposing narratives and polarization on the social platform. The study also displays a rather abstract contradiction of two conflicting collectives. The study does not necessarily define them as echo chambers, but does state that #TCOT is 'prominent in conservative echo chambers'.[126] Table 3 and 4 in the Ferguson study, however, show that within the opposing hashtags opinions supporting the tagged messages are so dominant that they make up 100% of the studied tweets. Therefore, these hashtags can be considered as

[124] K. Garimella and I. Weber, 'A Long-Term Analysis of Polarization on Twitter' (2017) <https://arxiv.org/abs/1703.02769>(7 November, 2017).
[125] R. Ray et. al. 'Ferguson and the death of Michael Brown on Twitter', appendix.
[126] R. Ray et. al. 'Ferguson and the death of Michael Brown on Twitter', p.1807.

echo chambers themselves. As explained is chapter 1, Twitter users can participate on three layers of communication: a micro, meso and macro level. Conflicting opinions reach opposing collective identities at a macro level under neutral hashtags and terms, as is shown in figure 8.

4.2 Focus of the Case Study

In this case study, the #BlackLivesMatter will be explored to see to what extent conflicting opinions appear within tweets using the hashtag themselves, to see if users advocating the BLM movement are being confronted with opposing views within their own hashtags. Next to this analysis, the case study will serve another purpose; it will explore the possibilities to analyze textual Twitter data with a modest set of text analysis tools. The data itself is merely textual, therefore the metadata has to be extracted from the text itself. By working with textual data, the study approaches the archive of the Library of Congres, which is text based too. The text analysis tools used, comes from the Master program Book and Digital Media in Leiden. Here they are mainly used to analyze literature and therefore longer texts. Therefore this study will also indicate to what extent a student, in a master within the humanities field, with limited and basic knowledge of textual analysis, can explore a data-driven platform such as Twitter.

4.3 Data

The data for this study was acquired via Twitter's REST API, with the help of a query written in Python. The query requested any tweet containing the #BlackLivesMatter. This was done during the period of 1 September until 7 October. This study eventually used the data from 15 September until 30 September in 2017. This resulted in a total of 129.690 tweets collected from this period.

Unlike the study focused on Ferguson, this data is not focused around a particular event. Because #BlackLivesMatter is one of the most popular hashtags, it is used on a daily basis. Most hashtags have the tendency to rapidly form and dissolve, and are therefore bound to a certain period. The #BlackLivesMatter, however, has become a platform Twitter users contribute to on a daily basis. It still peeks around topics regarding police violence against black people though, the matter that made the hashtag popular initially. Therefore it was not deemed necessary to focus on a certain period, or aftermath of an event. Regarding the question whether the hashtag operates as an echo chamber, it would be interesting to check the contributions made to #BlackLivesMatter in a relatively quiet period. This way the study can analyze whether recurring BLM related events keep the hashtags lively, or activist do by either fighting for justice or remembering past events. Apart from studying the contributors, this approach will also verify if analyses of the tweets can be sufficient to recapture a historic event, thereby testing Twitter as an informational source.

4.4 Methods and Results

With a term-documented matrix (tdm) script made with Perl, an overview of the 150 most frequent words used during the period was created.[127] The tdm file works with a list of stop words, in order to extract only relevant terms. With the help of the tdm file, the data can be approached top down. The most frequently used term can be used to capture the broader context of the period, after which less frequently used terms can serve to explore smaller details. As the files are merely textual, the number of times #BlackLivesMatters is mentioned is similar to the number of tweets.[128] In adjustment, some of the most common metadata, such as number of tweets and retweets, can be extracted with tdm from the files. After #BlackLivesMatter the most common terms are 'rt' and 'https', which are representing Twitter's main operators, retweets and link sharing. Therefore the first visualization is a line graph of the number of tweets, retweets and links, to determinate the number of unique tweets.

Table 1: Information Diffusion

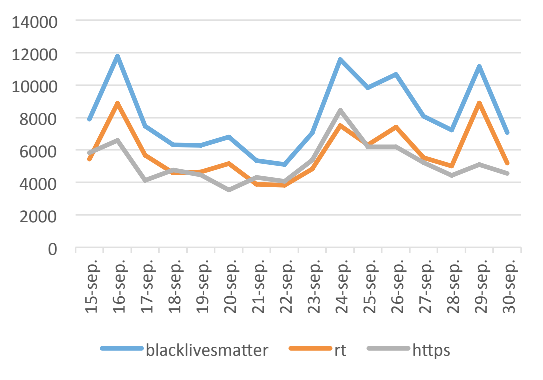


Table 1 reveals that the relative number of retweets is very constant. The number of retweets is always higher than the number of unique tweets, which is represented by the distance between the 'blacklivesmatter' and 'rt' line. Generally, daily about 2000 tweets are unique during the period, indicating that most of the data collection consist of 'top tweets': tweets that are retweeted frequently and appear in the top section of the #BlackLivesMatter.

---

[127] For the whole tdm, see the appendix on page 71.

[128] It is possible to use the hashtags multiple times in one tweet, but no examples of this have been spotted. It has no function to use the same hashtag twice and it takes up characters in an already limited space. Therefore it can be assumed that the number of times #BlackLivesMatter is similar to the number of tweets, or at least very near to it.

The ‘https’ line follows the ‘rt’ line, indicating that the most popular tweets most likely contain embedded content such as images, video’s or blogs.

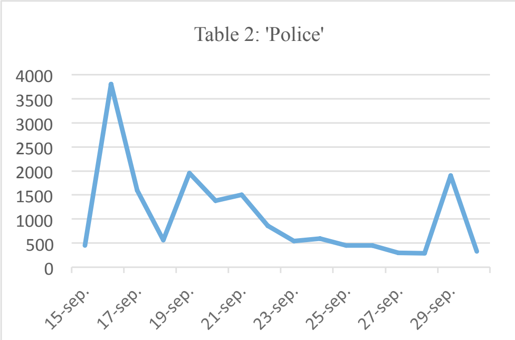


Table 2: 'Police'

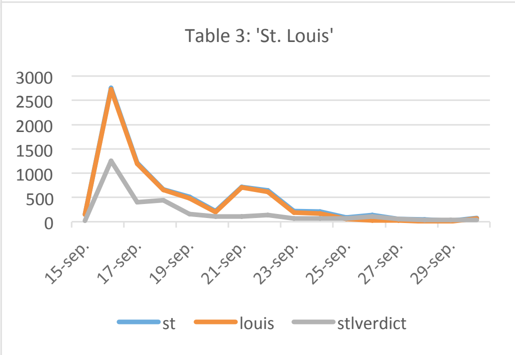


Table 3: 'St. Louis'

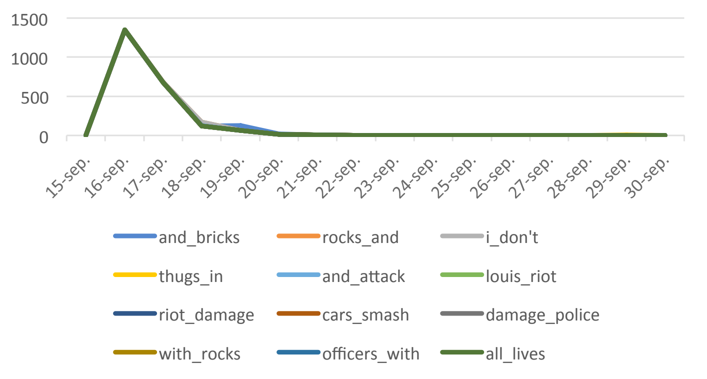
After these three terms that in fact represent metadata, the contextual tweets make their appearance in the tdm. ‘Police’ is the most present term overall. It is continuously a topic mentioned more than a hundred times, but it spikes around certain days. In the same manner there are terms that are present during the whole period, while other appear on a certain date. The terms present over the whole period can be used to analyze the general narrative present in #BlackLivesMatter, while the terms that spike at (a) certain day(s) can be helpful to determine the both influential events and/or tweets during this period.

Table 2 and 3 combined indicate an influencing event. During the whole period of 15 until 30 September, there were protests in St. Louis in reaction to the acquittal of former police officer Jason Stockley, who fatally shot a 24-year old black after a high speed chase in 2011.[129] This event is very similar to the Ferguson case of Mike Brown, therefore it can be expected that the police is criticized within #BlackLivesMatter. This hypothesis can be further explored with the help of two other Perl script. The first is the bigram script, that operates in the same way as the tdm script, except that it extract the terms that are frequently used combined. Therefore, it will spot top tweets more efficiently. To confirm that frequently used terms are indeed top tweets, concordace.pl[130] can be used to extract all sentences with, as

[129] T. O’Neil and M. Smith, ‘Former St. Louis Officer, Jason Stockley, Acquitted in Shooting of Black Driver’, *The New York Times,* 15 September, 2017 <https://www.nytimes.com/2017/09/15/us/jason-stockley-anthony-lamar-smith-st-louis-officer.html>(14 November, 2017).
[130] This tool can be downloaded from the BookandByte, ‘Digital Text and Data Processing – File Repository’ <http://bookandbyte.org/DTDP/index.php/file-repository/>(14 November, 2017).

demonstrated police, to get hold of a broader context of the top tweet.

Table 4: Retweet bigrams



All the terms in table 4 are related to one tweet, that is very dominant within the spike of activity during 15 until 19 September. This dominance is further demonstrated in the example

```
ugs in St. Louis riot, damage police cars, smash windows and attac
ars, smash windows and attack police officers with rocks and brick
ugs in St. Louis riot, damage police cars, smash windows and attac
ars, smash windows and attack police officers with rocks and brick
ugs in St. Louis riot, damage police cars, smash windows and attac
ars, smash windows and attack police officers with rocks and brick
ugs in St. Louis riot, damage police cars, smash windows and attac
ars, smash windows and attack police officers with rocks and brick
ugs in St. Louis riot, damage police cars, smash windows and attac
ars, smash windows and attack police officers with rocks and brick
ter Cowards Try To Intimidate Police and Get a Little Taste Of The
 smash windows and dance on a police car with a CHILD. #STLVerdict
ter Cowards Try To Intimidate Police and Get a Little Taste Of The
ugs in St. Louis riot, damage police cars, smash windows and attac
ars, smash windows and attack police officers with rocks and brick
```

Table 5: A fraction of the concordance.pl with the term ‘police.’

of the concordance about the term ‘police’, varying tweets appear only incidentally.

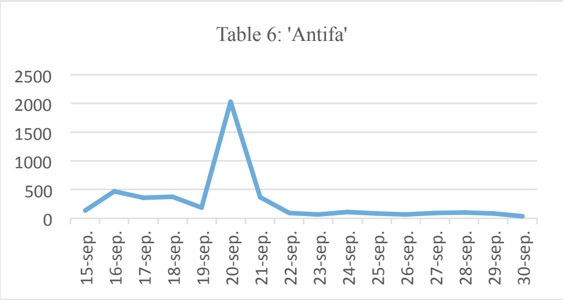
With the help of the actual text file of 16 September the actual tweet and user responsible for it can be extracted. As long as this user is still active, it can be traced with the help of Twitter or Google. In this case the spike was caused by the following tweet.

@_Makada_: #BlackLivesMatter thugs in St. Louis riot, damage police cars, smash windows and attack police officers with rocks and bricks. MSM IS SILENT[131]

This tweet is clearly not supportive of the BLM movement. With 'MSM' it directs at the mainstream media, arguing they remain silent considering the violence executed by BLM supporters. To add to the context, a video is added were supporters jump on the front window of a police car, breaking it. The user @_Makada_ currently has over 111 thousand followers and describes herself in her Twitter bio as a 'Nationalist, conservative, artist, writer and @American_Mirror contributor'.[132]

With the same approach that directed to this top tweet other influential users and their tweets can be extracted. Leading to the following top tweets during the whole analyzed period:

@PrisonPlanet: Antifa beats up Trump supporters. Trump supporters give #BlackLivesMatter a platform & let them speak. Big difference.[133]

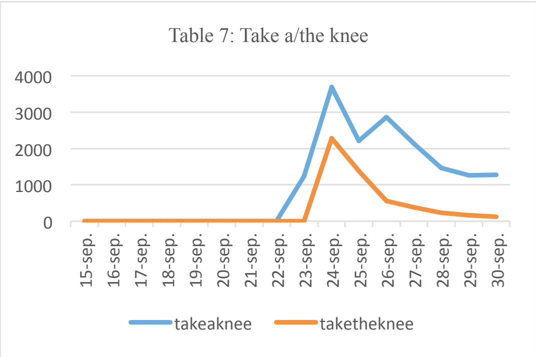


This one is from Paul Joseph Watson who has over 700 thousands followers and describes himself as 'Infowars editor-at-large. Classical liberal. Anti Alt-Right, anti Alt-Left.'.[134] Again, a video is added which shows how a BLM supporter gets invited on stage by Trump supporters. The video portrays a positive message, in which Trump supporters and BLM supporters unite, a video that itself was retweeted over 44 thousand time. Antifa, the anti-fascist, are a recurring term in the #BlackLivesMatter, as they are often related with the BLM-movement. Top tweets such as these indicate how part of the discussion is often concerned with distinguishing sides, while at the same time a unifying video is being spread.

[131] For the original tweet, see Twitter, '_Makada_' <https://twitter.com/_makada_/status/909073371207528448>(14 November, 2017).
[132] NB: Biographies can be changed over time by the users. https://twitter.com/_Makada_
[133] Twitter, 'PrisonPlanet' <https://twitter.com/PrisonPlanet/status/910462024030867457>(14 November, 2017).
[134] Twitter, 'PrisonPlanet' <https://twitter.com/PrisonPlanet >(14 November, 2017).

The bigrams indicate another top tweet from Watson.[135] On 26 September he tweeted the following thing:

> @Prisonplanet: Your daily reminder that the entire premise which underpins #TakeAKnee and #BlackLivesMatter has been debunked.[136]



Attached is an article from the Washington times titled 'No racial bias in police shootings, study by Harvard professor shows'.[137] The article brings evidence that opposes the view of the BLM movement. Therefore this tweet can be seen as unsupportive of the #BlackLivesMatter. In addition it is unsupportive of the #TakeAKnee, which is representative for NFL players kneeling during the national anthem following the NFL player Colin Kaepernick, who started this form of demonstration to protest against racial inequality and police brutality.[138] This hashtag became popular in #BlackLivesMatter, as shown by the visualization in table 7 from the tdm document.

Another opposing and dominant retweet in the data set comes from Mark Dice, who is a media-analyst with 290 thousand followers on Twitter and over a million on YouTube were he describes himself as someone who 'exposes liberal lunatics, celebrity scum, mainstream media manipulation' and is 'a social justice warrior psychos.'[139] His top tweet was formulated in the following way:

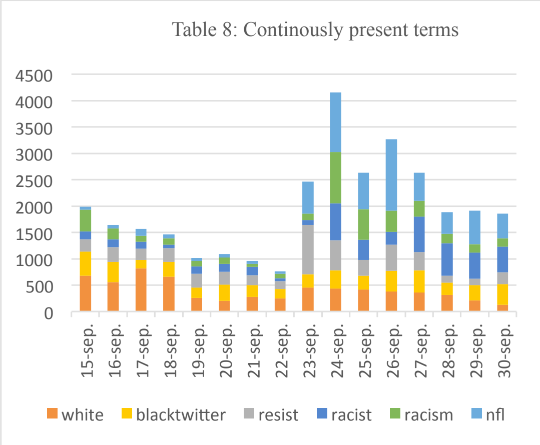
> @MarkDice: By saying Russia bought Facebook ads for #BlackLivesMatter to cause racial division, I guess CNN is finally admitting BLM is bad for America[140]

---

[135] See the appendix for the visibility of retweets in the bigram file on page 75.
[136] Twitter, 'PrisonPlanet' <https://twitter.com/prisonplanet/status/91272048720948838514 November, 2017).
[137] V. Richardson, 'No racial bias in police shootings, study by Harvard professor shows' <https://www.washingtontimes.com/news/2016/jul/11/no-racial-bias-police-shootings-study-harvard-prof/>(14 November, 2017)
[138] Know your Meme, '#TakeAKnee' <http://knowyourmeme.com/memes/events/takeaknee>(17 November, 2017).
[139] Youtube, 'User – Mark Dice' <https://www.youtube.com/user/MarkDice/about>(14 November, 2017).
[140] Twitter, 'Mark Dice'<https://twitter.com/markdice/status/913620317297303552?lang=en>(14 November, 2017).

This is again an example of how influential Twitter users spread a message attacking the BLM movement, in their own hashtag.

These top tweets that are dominant in the data set, making up for the majority of the dominant terms and bigrams. For the more general discussion regarding the BLM movement, terms in de tdm can be explored which are continuously present. A selection of these terms are displayed in table 8. They are generally present, but do not often exceed the limit of 1000 mentions.



Yet, when these terms are explored with the concordance tool, retweets appear to be dominant again. They can either be from news outlets, or influencers with a smaller range. One example is this tweets from HuffPost BlackVoices an account for ‘Black news, culture, entertainment and opinion’:[141]

@blackvoices: #TakeAKnee isn't about the flag. It's about America's racism.[142]

Attached is an image, proclaiming the following message:

> ‘Thinking the NFL players are ‘protesting the flag’ is like thinking Rosa Parks was protesting public transportation.’[143]

Tweets such as these have a couple of 100 retweets, in this case 307. With merely textual data and minimal tools it is hard to picture the conversation between non influential individuals. Tools such as the term-documented matrix, bigrams and concordance are suitable for picking out the most recurring tweets though. The negative effect of this though is in a study such this, it becomes rather complicated to take the focus away from these top tweets.

4.5 Case-study conclusion

As could be expected from the prospect of table 1, the data-set is highly influenced by retweets. Focusing on word frequency, the tools used are unable to extract contributions by non-influential Twitter users. Nonetheless the question whether #BlackLivesMatter acts as an

[141] Twitter, ‘Blackvoices’ <https://twitter.com/blackvoices>(14 November, 2017).
[142] Twitter, ‘Blackvoices’ <https://twitter.com/blackvoices/status/912455435302703109>(14 November, 2017).
[143] Ibidem.

echo chamber can partly be answered. While in the study regarding Ferguson supporting hashtags were dominant, the case study showed that opposing views are also very present in the period between 15 and 30 September. With the help of a large follower base, online influencers are able dominate the top tweets in the hashtag. Thereby they can partly change the dominant view within #BlackLivesMatter during a period. It is complicated to determine however if this is perceived this way, because Twitter's unique look and feel is lost by the textual format. Twitter's interface considering hashtags offers the 'top' and 'latest' story function. In the top feed, the tweets extracted in the case study will appear. Most of these retweets however, do not add a comment, and as they remain blank, they do not appear in the 'latest' story function, because they do not contain #BlackLivesMatter. The 'latest' feed therefore gives a whole other perception of #BlackLivesMatter, which is the conversation between individuals instead of influencers.

Therefore this case study can only conclude that within the 'top' feed opposing opinions exist and #BlackLivesMatter does not operate as an echo chamber. To explore the 'latest' feed it would be necessary to remove all blank retweets and order the individual contributions to analyze this conversation. For this two different options could be operable. Firstly, two types of request should be sent to the REST API at the same time. One including, and one excluding retweets. Secondly, advanced tools could be used to remove the retweets from the textual format.

The case study has shown that I, as a master student within the humanities, with limited tools and assistance, was able to retrieve a data set, but unable to work around the top tweet bias. Yet, with the help of sufficient knowledge from Twitter studies regarding the BLM movement, I have been able to spot a change. The once so uniform #BlackLivesMatter is no longer an echo chamber that voices an uniform message, the once so stable narrative has dissolved and is now being 'invaded' by opposing opinions.

General Discussion: Too few or too many data?

Popular hashtags and terms such as #BlackLivesMatter and Ferguson provide scholars with an abundance of data to study racial issues in the United States. They offer opportunities to explore social cohesion and diversion: where it starts, how it evolves and - perhaps in the future - where it will end. Yet, even the study regarding Ferguson, operating with millions of tweets maintains a select focus of the whole BLM movement, by sticking to popular hashtags around the term Ferguson. Although the case study had a more experimental approach and smaller data set, it did reveal important assets of Twitter's ecology. It showed how online influencers can put their mark on a debate, empowered by a large audience. These top tweets are often accompanied by news items, blogs, videos and photos. Therefore, Twitter itself operates in the far more extensive environment called the 'media ecology'. Being a medium, Twitter is 'a technology within which a culture grows; that is to say, it gives form to a culture's politics, social organization, and habitual ways of thinking'.[144] Twitter influence on the BLM movement therefore cannot be denied, but at the same time it must be acknowledged that Twitter operates as a wheel within a larger media system.

Does this mean that to fully understand the BLM movement on Twitter, scholars must capture every single tweet concerned with BLM and every multimedia item attached to it? This would at the first place be very problematic, as chapter 2 has explained how complicated and costly it can be to collect these quantities of data. In the second place, scholars need to consider if such a holistic approach of data is deemed necessary. 'Big data is not necessarily better data' is the message by Christine Borgman, the distinguished professor in information studies and author of *Big Data, Little Data, No Data: Scholarship in the Networked World* in which she further explains her message as followed:

> 'The farther the observer is from the point of origin, the more difficult it can be to determine what those observations mean- how they were collected; how they were handled, reduced, and transformed; and with what purposes in mind. Scholars often prefer smaller amounts of data that they can inspect closely. When data are undiscovered or undiscoverable, scholars may have no data.'[145]

[144] N. Postman, 'The Humanism of Media Ecology', *Proceeding of the Media Ecology Association, Volume 1* (2000) <http://www.media-ecology.org/publications/MEA_proceedings/v1/postman01.pdf>(14 November, 2017).
[145] C.L. Borgman, *Big data, little data, no data : scholarship in the networked world* (Cambridge: The MIT Press, 2015), p. xvii.

The true challenge regarding data collection might therefore not be to collect bulky data sets, but to collect carefully demarcated data sets. The study regarding Ferguson, but also the case study, use demarcated data-sets of the broader BLM movement. Thereby, they allow interested scholars to follow the process of data collection and handling. The results are collective identities, patterns and narratives in the form of tables or visualizations, which at their turn can be used to ask different questions. New studies can borrow the method from the Ferguson study to explore collective identities concerning economical or religious issues in either the United States or other countries. Hereby, they can analyze small systems operating in a larger Twitter ecology and even larger media ecology, to create more knowledge and eventually explore different kinds of data, questions, patterns and insights.[146]

Although scholars should be wary of aiming for large and undiscoverable data-set, Twitter data has also restrictions that have to be taken into consideration when studying the platform. What cannot be retrieved yet for scholars, is data reflecting how users spend their time on Twitter. This might be a luxury element of online data, but it can be of great value for scholars. For Twitter it would be useful to know how much time its user spend within different communication layers. Within their interface users can explore their own feed, the meso level where messages are displayed of people they follow. When users explore hashtags at a macro level, they can divide their time between the 'top' and 'latest' feed. These user statistics might be very useful to see were users spend their time the most, and therefore to decide to what extent they consume information from their filter bubble (meso level) or are confronted with opposing views at a macro level. This is especially useful to contribute to the debate of polarizing effect of social media. This debate tends to focus at biased information users perceive, while as shown in the case study, Twitter is a platform where opposing ideas clash in public conversation. Offending people, appending blunt opinions crude jokes and sexist commentary by a small group has proven to be effective to reach a large audience.[147] The role of the effect of provocation online is something that should be studied further, especially in the global informational warfare, were countries such as Russia use trolls to influence countries by creating division among people online.[148]

[146] L. Manovich, 'Trending: The Promises and the Challenges of Big Social Data' in M.K. Gold, *Debates in the Digital Humanities* (Minneapolis: University of Minnesota Press, 2012).
[147] A. Leavitt, 'From #FollowFriday to YOLO: Exploring the Cultural Salience of Twitter Memes' in in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014), p. 145.
[148] See for example this article, that describes how social media accounts run by trolls fueled activism. S. Levin, O.Solong and S. Walker, ''Our pain for their gain': the American activists manipulated by Russian trolls.', *The*

Another factor that influences the online conversation on Twitter are the demographics and personal traits of people who use Twitter. One of the most promising applications of Twitter data is that scholars can take surveys without asking questions. This way they can evade the response bias, the tendency for participants to give answers that are not accurate or truthful. This survey approach was used in ea. *Personality and Politics*, where personality types were extracted from personal feeds. In the results certain personality types were overrepresented. This does not mean, however, that the methods used did not work correctly, some personalities feel more urge to speak than others. One of the biggest questions regarding Twitter users is why certain people have the tendency to speak out, while others remain silent. Answers for this question might be found by exploring personalities and demographics. Considering demographics, it is known for the United States Twitter users do not accurately represent the population. Young adults (18-29 year) are overrepresented , and so are black people and Hispanics.[149] Among white people 6% say that most of the posts they see are about race, compared to 24% among black people.[150]

These statistics are gathered by taking surveys among users. They create valuable information, and provide contrast between what scholars perceive from the data and what Twitter users perceive and experience the online discussions. So apart from focusing on vast amounts of data, demarcated data or a collections of just 100 tweets, it is still necessary to question groups of users or interview individuals regarding to their Twitter experience. Both big, quantitative, data and small, qualitative data can be used together to provide insights in the dynamics that shape the public conversation on Twitter. This is very well demonstrated in the *Gift of the Gab* study, were quantitative analysis is used to create cluster visualizations of the Favstar network; a widely popular network of Twitter users in Germany. Here they support there quantitative analysis with personal interviews with Favstar, causing further insight in the 'gift economy' of retweet and like cartels.[151] Therefore the real challenge is not the volume or type of the data set, but uniting all kinds of data sets into one structured narrative.

Guardian, 21 October, 2017 <https://www.theguardian.com/world/2017/oct/21/russia-social-media-activism-blacktivist>(14 November, 2017).

[149] M. Duggan, 'Mobile Messaging and Social Media 2015', *Pew Research Center,* 19 August, 2015 <http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/>(14 November, 2017).

[150] M. Anderson and P. Hitlin, 'Social Media Conversations About Race', *Pew Research Center*, August 2016 <http://www.pewinternet.org/2016/08/15/the-hashtag-blacklivesmatter-emerges-social-activism-on-twitter/>(14 November, 2017).

[151] J. Paßmann, 'The Gift of the Gab', p. 341-343.

Conclusion

The goal of this thesis was to explore the *potential* and *challenges* of Twitter data for the humanities. First of all, evaluating Twitter's structure showed that scholars need to get insight into the dynamics of Twitter. In the first ten years of the platform's existence, communication has been subject to changes coming from both Twitter's community and the company itself. Twitter's transforming structure will make it hard to recapture its authentic feel as the platform keeps evolving, especially when the data is preserved -as for example in the Library of Congress- in a textual form. Twitter's structure, however, is simplistic and stayed close to its original. Therefore, it will be easy for humanities scholars to understand the platform and develop research questions for it.

The challenges considering data retrieval are of a computational nature and will therefore form barriers for most of the scholars in the humanities. Apart from that they encounter restrictive, policy-orientated measures, degrading the amount of available data. On the short term, the prospect for sufficient data retrieval, is that humanities scholars are either dependent on computational skilled assistants and/or financial needs to acquire the Firehose stream or any historical data. Therefore, it is desirable that other public institutions assist or take over the Library of Congress' Twitter archive, to provide scholars with free, structured, accessible data. On the long term, disciplines within humanities can either choose to invest more time in developing, and/or learning about, computational skills and methods, possibly by prioritizing the digital over the traditional humanities.

Despite the list of challenges, Twitter provides plenty of opportunity to study aspects of human society and culture. As a wide variety of scholarly disciplines is engaged in Twitter studies, there are plenty studies providing data and visualizations to *extract* knowledge from. Here, the only challenge for humanities scholars is to become known with Twitters dynamics, data, visualizations and develop the literacy to understand it. Humanists, however, also *apply* their own research questions and concepts to Twitter data. This can create unique opportunities within various disciplines within humanities. Historians can study the rise of demonstrations, civil movements, the first reaction to critical news and commemoration. Linguists can explore language use on a large scale, types of discourse and how they implicitly or implicitly deliberately cause division. With the help of data, political historians or scientist can carefully follow the narrative swings of politicians and its public reception. Philosophers, finally, can discuss the role of global communication and information diffusion in human society on a broader level.

These are all examples from this thesis, that evaluated a relatively young field of study, that has needs to develop further to explore its full potential. Social media platforms such as Twitter, but also Facebook and YouTube, play an increasing role in what we read, and therefore how we perceive the world around us. In the following years, social media will claim more of our digital attention, money and therefore our mind.[152] Understanding this global phenomenon and its impacts on human society is perhaps the biggest challenge for humanists in history so far. This thesis has shown that every platform needs to be understood in its unique way, that both trends in large data sets, details in small datasets, but also close reading and interviewing are necessary tools to capture the bigger picture. Therefore the humanities will have to adapt and extend their traditional form of reading to challenge these big issues coming with social media. This way 140 character long 'snippets' can be used to write 'stories', which help us to understand our role as human in a mediatized world.

---

[152] To get an indication of how time will be spend on media in the upcoming years, see: D. Clark, 'What next for Tech and Media in 2017', *The Wall Street Journal,* October 2016 <https://www.wsj.com/articles/activates-michael-wolf-predicts-whats-next-for-tech-and-media-in-2017-1477436031#>(19 November, 2017).

Bibliography

Secondary sources

Anon., 'Twitter tweets are 40% babble', *BBC News,* 17 August, 2009 <http://news.bbc.co.uk/1/hi/technology/8204842.stm> (18 September 2017).

Adnan, M., A., Leak and P. Longley, 'A Geocomputational analysis of Twitter activity around different world cities', *Geo-spatial Information Science,* (2014), p. 1-8.

Anderson, M., and P. Hitlin, 'Social Media Conversations About Race', *Pew Research Center*, August 2016 <http://www.pewinternet.org/2016/08/15/the-hashtag-blacklivesmatter-emerges-social-activism-on-twitter/>(25 October, 2017).

Beaumont, C., 'New York plane crash: Twitter: breaks the news, again.' *The Telegraph,* 16 January, 2009 <http://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html> (18 September, 2017).

Blonk, M., V. Buss, M. Scherf, J. Voorn and P. Vliegenthart, ,'Analysing tweets of Trump and Wilders', *Digital Methods Initiative, Winter School 2017*, (2017). <https://wiki.digitalmethods.net/Dmi/WinterSchool2017TrumpWildersTweetAnalysis>(18 October, 2017).

Boxell, L., M. Gentzkow and J. Shapiro, 'Is the Internet Causing Political Polarization? Evidence from Demographics', *NBER Working Paper Series*, 2017 <https://www.brown.edu/Research/Shapiro/pdfs/age-polars.pdf >(7 November, 2017).

Borgman, C.L., *Big data, little data, no data : scholarship in the networked world* (Cambridge: The MIT Press, 2015).

Bruns, A., and H. Moe, 'Structural Layers of Communication on Twitter', in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Bruns, A., and K. Weller, 'Twitter as a First Draft of the Present – and the Challenges of Preserving It for the Future', *Preceedings of ACM Web Science Conference,* (2016) <http://dx.doi.org/10.1145/2908131.2908174 >(7 November, 2017).

Clark, D.,'What next for Tech and Media in 2017', *The Wall Street Journal,* October 2016 <https://www.wsj.com/articles/activates-michael-wolf-predicts-whats-next-for-tech-and-media-in-2017-1477436031#>(19 November, 2017).

Clavert, F,, '#ww1. The Great War on Twitter', *Digital Humanities Congress 2017 Abstracts*, (2017).

Chen, E., 'Soda vs. pop with Twitter' <http://blog.echen.me/2012/07/06/soda-vs-pop-with-twitter/>(14 November, 2017).

Cobb, J., ‘The Matter of Black Lives’, *The New Yorker,* 14 March 2016 <https://www.newyorker.com/magazine/2016/03/14/where-is-black-lives-matter-headed>(25 October, 2017).

Collins, K., ‘The 10th anniversary of the hashtag is a reminder that Twitter’s best features came from outside the company’, *Quartz*, 24 August, 2017 <https://qz.com/1060789/the-10th-anniversary-of-the-hashtag-is-a-reminder-that-twitters-best-features-came-from-outside-the-company/> (18 September, 2017).

Duggan, M., ‘Mobile Messaging and Social Media 2015’, *Pew Research Center,* 19 August, 2015 <http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/>(14 November, 2017).

Edwards, J., ‘Leaked Twitter API data shows the number of tweets is in serious decline’, *Business Insider UK,* 2 February, 2016 <http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2?international=true&r=UK&IR=T>(10 October, 2017).

Garimella, K., and I. Weber, ‘A Long-Term Analysis of Polarization on Twitter’ (2017) <https://arxiv.org/abs/1703.02769>(7 November, 2017).

Grandjean, M., ‘A social network analysis of Twitter: Mapping the digital humanities community’, *Cogent Arts & Humanities, 3*(1), Cogent Arts & Humanities, Dec 2016, Vol.3(1) (2016).

Grifantini, K., ‘The Evolution of Retweeting’, *Technology Review,* 26 August, 2009 <https://www.technologyreview.com/s/415043/the-evolution-of-retweeting/> (18 September, 2017).

Halavais, A.,‘Structure of Twitter: Social and Technical’, in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Hepp, A., & F. Krotz, *Mediatized worlds : Culture and society in a media age.* (Basingstoke, Hampshire: Palgrave Macmillan, 2014).

Huffpost, ‘All Lives Matter’ <https://www.huffingtonpost.com/topic/all-lives-matter>(25 October, 2017).

Hockey, S., ’The History of Humanities Computing’ in S. Schreibman, R. Siemens and J. Unsworth, *A Companion to Digital Humanities* retrieved from <http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-2-1&toc.depth=1&toc.id=ss1-2-1&brand=default> (18 September, 2017).

Juola, P., and S. Vinsick, ‘Personality and Politics: Myers-Briggs Personality Types on Twitter in the US 2016 Presidential Election’, *Digital Humanities Congress 2017 Abstracts*, (2017).

Ingram, M., ‘Drinking from the Twitter firehose: I love the stream but I need more filters and bridges’, *Gigaom,* 9 January, 2014 <https://gigaom.com/2014/01/09/drinking-from-the-twitter-firehose-i-love-the-stream-but-i-need-more-filters-and-bridges/> (18 September, 2017).

Khomani, N., ‘#MeToo: how a hashtag became a rallying cry against sexual harassment’, *The Guardian,* 20 October, 2017 <https://www.theguardian.com/world/2017/oct/20/women-worldwide-use-hashtag-metoo-against-sexual-harassment>(14 November, 2017).

Koh, Y.‘Twitter Paid $134 Million for Data Partner Gnip’, *The Wall Street Journal,* 11 August, 2014 <https://blogs.wsj.com/digits/2014/08/11/twitter-paid-134-million-for-data-partner-gnip/> (18 September 2018).

Kreis, R.,‘#refugeesnotwelcome: Anti-refugee discourse on Twitter’, *Discourse & Communication,* Vol. 11 (5), 2017.

Krikorian, R., ‘Introducing Twitter Data Grants’, *Twitter Blog*, 5 February 2014 <https://blog.twitter.com/engineering/en_us/a/2014/introducing-twitter-data-grants.html> (18 September, 2017).

Krikorian, R., ‘Twitter #DataGrants selections’, *Twitter Blog*, 17 April, 2014 <https://blog.twitter.com/engineering/en_us/a/2014/twitter-datagrants-selections.html> (18 september, 2017).

Krikorian, R., ‘New Tweets per second record, and how!’ *Twitter Blog,* 16 August 2013 <https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html> (18 September, 2017).

Lanier, J., *You Are Not A Gadget: A Manifesto* (London: Penguin, 2010).

Leetaru, K., S. Wang, G. Cao, A. Padmanabhan, A. & E. Shook, ‘Mapping the global Twitter heartbeat: The geography of Twitter’, *First Monday, 18*(5) (2013) <http://firstmonday.org/article/view/4366/3654> (18 October, 2017).

Lindhout, S., ‘Onderzoek: Duitse kranten namen Merkels ‘Wilkommenskultur’ vrijwel kritiekloos over’, *De Volkskrant,* 24 July 2017 <https://www.volkskrant.nl/buitenland/onderzoek-duitse-kranten-namen-merkels-wilkommenskultur-vrijwel-kritiekloos-over~a4507954/>(7 Novermber, 2017).

Linn, J., ‘Studying rapidly evolving user interests’, *Twitter Blog,* 4 June 2012 <https://blog.twitter.com/engineering/en_us/a/2012/studying-rapidly-evolving-user-interests.html>(17 October, 2017).

McGill, A., ‘Can Twitter Fit Inside the Library of Congress?’, *The Atlantean*, 4 August, 2016 <https://www.theatlantic.com/technology/archive/2016/08/can-twitter-fit-inside-the-library-of-congress/494339/> (18 September, 2017).

Morstatteret F., al., ‘Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose’, (2013).

Norman., J.,‘Roberto Busa & IBM adapt Punched Card Tabulating Sort Words in a Literary Text: The Origins of Humanities Computing’, *HistoryofInformation.com* <http://www.historyofinformation.com/expanded.php?id=2321> (18 September, 2017).

Pariser, E., *The Filter Bubble : what the internet is hiding from you* (London, 2011).

Paßmann, J., T. Boeschoten and M.T. Schäfer, ‘The Gift of the Gab: Retweet Cartels and Gift Economies on Twitter’ in in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Puschmann, C., and Jean Burgess, ‘The Politics of Twitter Data’ in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Priego, E., ‘Father Roberto Busa: One academic’s impact on HE and my career’, *The Guardian,* 12 August, 2011 <https://www.theguardian.com/higher-education-network/blog/2011/aug/12/father-roberto-busa-academic-impact> (18 September, 2017).

Ray, R., M. Brown, N. Fraistat & E. Summers, ‘Ferguson and the death of Michael Brown on Twitter: #BlackLicesMatter, #TCOT, and the evolution of collective identities’, *Ethnic and Racial* Studies, 40: 11 (2017).

Raymond, M., ‘Twitter Donates Entire Tweet Archive to Library of Congress’, *Library of Congress*, 15 April, 2010 <https://www.loc.gov/item/prn-10-081/> (18 September, 2017).

Richardson, V., ‘No racial bias in police shootings, study by Harvard professor shows’ <https://www.washingtontimes.com/news/2016/jul/11/no-racial-bias-police-shootings-study-harvard-prof/>(14 November, 2017).

Rogers, R., ‘Debanalising Twitter: The Transformation of an Object of Study’, in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Rosen, A., and I. Ihara, ‘Giving you more character to express youself’, *Twitter Blog,* 26 September, 2017 <https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html>(14 November, 2017).

Ruiz Soler, J., ‘Twitter Research for Social Scientists: A Brief Introduction to the Benefits, Limitations and Tools for Analysing Twitter Data.’ *Dígitos: Revista De Comunicación Digital*, no. 3 (2017).

Sarno, D., ‘Twitter creator Jack Dorsey illuminates the site’s founding document. Part I’, *Los Angeles Times*, 18 February, 2009 <http://latimesblogs.latimes.com/technology/2009/02/twitter-creator.html> (18 September, 2017).

Sarno, D., ‘Jack Dorsey on the Twitter ecosystem, journalism and how to reduce reply spam. Part II’, *Los Angeles Times,* 19 February, 2017 <http://latimesblogs.latimes.com/technology/2009/02/jack-dorsey-on.html> (18 September, 2017).

Stone, B., ‘Twitter search for everyone!’, *Twitter Blog,* 30 April, 2009 <https://blog.twitter.com/official/en_us/a/2009/twitter-search-for-everyone.html> (18 September, 2017).

Tate, R., ‘Twitter’s new prompt: A linguist weighs in’, *Gawker,* 19 November, 2009 <http://gawker.com/5408768/twitters-new-prompt-a-linguist-weighs-in> (18 September, 2017).

Twitter Engineering, ‘200 million Tweets per day’, *Twitter Blog,* 30 June, 2011 <https://blog.twitter.com/official/en_us/a/2011/200-million-tweets-per-day.html> (18 September, 2017).

Valentino-De Vries, J., ‘Twitter Buys TweetDeck’, *The Wall Street Journal,* 25 May, 2011 <https://blogs.wsj.com/digits/2011/05/25/twitter-buys-tweetdeck/> (18 September 2018).

Wang, Y., J. Callan & B. Zheng, ‘Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API’ *ACM Transactions on the Web (TWEB)*, 9(3), (2015).

Warren, C., 'Twitter's API Update Cuts Off Oxygen to Third-Party Clients', *Mashable,* 16 August, 2012.

Weller, K., A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Wilken, R., 'Twitter and Geographical Location' in K. Weller, A. Bruns, J. Burgess, M. Mahrt and C. Puschmann, *Twitter and Society* (New York, Peter Lang Publishing, 2014).

Wodak, R. and S. Boukala 'European identities and the revival of nationalism in the European Union: A discourse-historical approach', (2015b) .

Wolny, W., 'Knowledge Gained from Twitter Data', *Annals of Computer Science and Information Systems,* 8, pp. 1133-1336.

Yin, J., A., Soliman, D., Yin, S., Wang, 'Depicting urban boundaries from a mobility network of spatial interactions: A case study of Great Britain with geo-located Twitter data.' *International Journal of Geographical Information Science, 31*(7), (2017), pp 1219-1313.

M. Zimmer and N.J. Proferes, 'A topology of Twitter research: disciplines, methods, and ethics.' *Aslib Journal of Information Management*, 66(3), 2014, pp.250–261.

Zimmer, M., 'The Twitter Archive at the Library of Congress: Challenges for information practice and information policy', *First Monday, 20* (7), (2015) <http://firstmonday.org/article/view/5619/4653> (18 September, 2017).

Websites and Digital Resources

4Humanities, 'What are the Humanities' <http://4humanities.org/2014/12/what-are-the-humanities/>(10 October, 2017).

American Library Association, 'Ethics' <http://www.ala.org/tools/ethics> (7 November, 2017).

Gnip, 'Historical' <https://gnip.com/historical/> (18 September, 2017).

Internet Live Satistics, 'Twitter Usage Statistics'<http://www.internetlivestats.com/twitter-statistics/>(17 November, 2017).

Know your Meme, '#TakeAKnee' <http://knowyourmeme.com/memes/events/takeaknee>(17 November, 2017).

Library of Congress, 'Update on the Twitter Archive At the Library of Congress', (2013) <https://www.loc.gov/static/managedcontent/uploads/sites/6/2017/02/twitter_report_2013jan.pdf> (18 September, 2017).

OED Online, 'humanity,n.' <http://www.oed.com.ezproxy.leidenuniv.nl:2048/view/Entry/89280?redirectedFrom=humanities#eid311537170>(15 November 2017).

Stanford Humanities Center, 'What are the Humanities?' <http://shc.stanford.edu/what-are-the-humanities> (18 September 2017).

Stanford University, 'Mapping the Republic of Letters' <http://republicofletters.stanford.edu/>(18 October, 2017).

Twitter, 'Company#about' <https://about.twitter.com/nl/company#about> (18 September, 2017).

Twitter Developer Documentation, 'GET Statuses/Lookup' <https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup> (25 October, 2017).

Twitter Development Documentation, 'Rate Limits: Chart' <https://dev.twitter.com/rest/public/rate-limits> (18 September, 2017).

Twitter Development Documentation, 'The Search API' <https://dev.twitter.com/rest/public/search> (18 September, 2017).

Twitter Files, 'Selected Company Metrics and Financials', <http://files.shareholder.com/downloads/AMDA-2F526X/5225764351x0x951002/FCE28680-E74E-4349-A11C-4B86BBABFB26/Q217_Selected_Company_Metrics_and_Financials.pd> (10 October, 2017).

Twitter Interactive, 'Philippines'<http://twitter.github.io/interactive/philippines/>(19 November, 2017).

Twitter, 'Twitter Interactive'<https://interactive.twitter.com/>(17 October, 2017).

Twitter, 'Terms of Service'<https://twitter.com/en/tos> (18 September, 2017).

Twitter, 'Twitter Support'<https://twitter.com/twittersupport/status/442433903546994688> (18 September 2017).

Twitter Support, 'Our Policy' <https://support.twitter.com/articles/20175050> (18 September, 2017).

Twitter Support, 'Privacy Policy' < https://support.twitter.com/articles/20174657> (18 September, 2017).

Voyant Tools <https://voyant-tools.org/> (18 September, 2017).

Wikipedia, ‘Confirmation Bias’ <https://en.wikipedia.org/wiki/Confirmation_bias>(7 November, 2017).

YouTube, ‘User – Mark Dice’ <https://www.youtube.com/user/MarkDice/about>(14 November, 2017).

Zephoria, ‘Strategic Insights’ <https://zephoria.com/top-15-valuable-facebook-statistics/> (18 September, 2018).

Tweets (& Users)

Twitter, ‘Blackvoices’ <https://twitter.com/blackvoices/status/912455435302703109>(14 November, 2017).

Twitter, ‘_Makada_’ <https://twitter.com/_makada_/status/909073371207528448>(14 November, 2017).

Twitter, ‘Mark Dice’<https://twitter.com/markdice/status/913620317297303552?lang=en>(14 November, 2017).

Twitter, ‘PrisonPlanet’ <https://twitter.com/PrisonPlanet/status/910462024030867457>(14 November, 2017).

Twitter, ‘PrisonPlanet’ <https://twitter.com/PrisonPlanet/status/910462024030867457>(14 November, 2017).

Twitter, ‘PrisonPlanet’ <https://twitter.com/prisonplanet/status/912720487209488385>(14 November, 2017).

Appendix

police	takeaknee	black	amp	blm	st	louis	trump	nfl	white	lives	protest	people	don't	officer	taketheknee	resist	matter
454	0	993	309	561	151	149	74	62	681	153	121	186	241	1097	0	229	129
3807	3	1152	725	642	2757	2730	68	69	561	294	483	450	229	586	0	279	231
1601	3	1022	351	455	1212	1190	104	127	814	249	325	188	250	137	0	216	201
569	1	791	271	673	667	652	63	75	656	199	132	402	120	48	0	257	165
1954	1	562	287	433	516	485	74	62	257	106	232	217	108	813	0	260	107
1377	2	535	2042	724	219	204	3730	67	198	184	56	156	72	99	0	243	126
1502	2	1059	382	340	719	710	660	56	278	91	86	574	91	123	0	189	79
862	1	559	223	313	642	615	89	52	247	132	76	572	194	31	0	150	96
543	1239	1203	406	369	215	187	260	605	453	104	126	536	129	11	6	929	75
597	3685	1573	855	545	213	172	491	1126	434	653	661	389	291	75	2273	570	584
455	2216	887	891	448	90	58	475	692	417	351	299	293	194	20	1381	304	316
454	2859	949	816	643	135	31	276	1356	379	262	641	617	509	27	552	490	210
303	2123	708	574	646	52	30	188	531	356	704	277	262	726	10	377	345	693
289	1472	891	490	277	47	5	171	407	312	1049	134	236	601	120	231	128	666
1900	1257	963	760	2160	20	10	100	634	211	846	1471	194	802	1794	158	116	706
332	1274	536	742	545	77	63	208	470	125	517	365	173	766	237	124	225	484

bap	supporters	prisonplanet	america	racism	sued	cnn	movement	justice	support	attack	does	facebook	russia	markdice	stlverdict	know	judge
317	6	0	364	410	0	23	46	536	48	139	51	5	5	0	20	977	50
254	8	2	143	208	0	92	81	518	116	2260	574	1	0	2	1255	321	81
233	1	3	452	118	0	51	153	274	95	782	523	3	1	4	407	134	25
255	7	2	83	120	0	38	76	241	64	157	74	7	1	1	440	60	2
245	10	5	60	97	0	154	46	189	46	93	33	6	0	1	161	54	0
320	3685	1560	49	116	0	37	38	124	56	16	23	22	1	0	103	62	0
234	610	191	48	58	0	619	68	121	44	9	23	8	5	0	106	31	3
210	51	12	35	82	0	207	114	92	44	3	27	3	2	0	135	39	1
295	19	2	139	126	0	62	98	197	165	13	29	1	5	1	68	68	1
413	110	7	886	980	0	66	339	255	356	8	55	6	4	91	71	145	1
341	19	26	406	576	0	182	195	205	219	6	45	246	214	204	67	134	0
420	14	2098	465	402	2	99	122	146	249	205	129	556	158	18	106	82	0
539	8	563	182	308	0	154	82	118	733	52	616	214	40	11	58	85	1
112	8	54	100	181	779	126	805	102	631	8	519	421	186	14	34	95	824
244	7	16	155	159	2639	1803	1095	186	482	14	552	1831	2519	2559	40	404	1473
147	13	10	593	161	657	259	452	484	437	1	446	246	322	325	35	357	547

TDM

title	thugs	deray	cops	right	protesting	kaepernick	officers	ads	let	speak	windows	killed	makada	bricks	man	aclu	make	riot	racial	like	smash	realdonaldtrump
15-sep	16	13	157	285	858	107	43	1	25	11	2	1404	1	0	268	1237	48	67	14	151	1	15
16-sep	1771	47	496	456	218	65	1524	1	25	44	1741	655	1728	1685	132	200	82	1494	23	270	1649	56
17-sep	839	27	131	66	78	111	716	0	24	36	732	109	740	711	155	46	496	691	19	150	686	52
18-sep	158	10	117	35	68	51	160	0	17	16	133	67	127	187	52	33	73	166	14	99	119	24
19-sep	77	15	627	27	24	33	207	0	15	23	67	42	68	86	819	20	57	69	10	85	64	45
20-sep	16	9	198	70	11	44	48	0	1578	1565	24	59	12	11	114	13	56	23	8	74	12	24
21-sep	14	9	97	29	43	25	13	0	200	197	7	27	6	7	204	391	411	13	6	79	6	17
22-sep	4	8	105	15	33	14	9	2	19	44	6	21	3	2	82	480	479	10	9	75	2	10
23-sep	16	14	80	548	55	365	13	2	30	29	7	21	4	3	82	102	149	13	17	102	3	302
24-sep	18	32	85	361	112	938	9	0	75	45	2	43	1	1	79	29	128	18	142	473	3	836
25-sep	13	18	92	231	80	359	24	242	104	108	2	120	3	3	75	3	87	6	37	199	2	532
26-sep	6	45	150	252	97	259	73	164	431	103	3	86	3	1	132	43	82	19	73	204	1	239
27-sep	7	20	109	83	71	208	28	77	111	136	3	28	4	3	295	22	82	12	54	234	4	137
28-sep	7	526	58	62	374	110	26	333	73	38	0	25	1	0	87	9	58	5	47	153	1	72
29-sep	10	1952	58	162	491	135	16	1833	21	146	0	11	3	0	58	17	134	2	1712	169	0	40
30-sep	5	212	387	264	322	93	8	218	19	191	0	9	0	1	50	8	191	0	401	67	0	98

title	mean	saying	stump	larryelder	cause	damage	say	cars	video	rocks	flag	men	stl	bought	protesters	anthonylamarsmith	violence	guess	rights	wearing	actually	big
15-sep	17	77	30	8	10	0	196	1	68	0	5	23	272	1	55	565	242	7	854	0	13	6
16-sep	51	98	780	435	141	1357	115	1360	234	1351	183	50	389	0	625	767	176	15	175	4	14	74
17-sep	7	47	176	430	163	679	88	678	503	686	22	68	291	0	119	202	63	4	67	6	8	23
18-sep	15	12	311	50	27	217	75	120	74	172	2	17	333	0	313	144	30	12	42	8	58	9
19-sep	10	14	395	15	12	66	145	130	56	64	19	17	78	1	142	242	34	6	25	994	4	13
20-sep	13	12	529	10	7	16	51	19	297	13	30	41	43	0	364	56	52	12	19	715	7	1569
21-sep	5	14	29	597	6	6	624	7	257	6	15	604	43	2	317	18	612	1	48	92	596	196
22-sep	5	31	4	170	8	3	197	4	62	2	8	208	112	0	45	36	214	25	38	40	180	21
23-sep	14	27	4	282	14	4	90	5	28	5	83	719	101	2	27	6	290	8	37	46	39	17
24-sep	47	32	1	340	25	3	254	1	47	1	578	283	212	3	24	11	112	29	207	66	172	49
25-sep	18	23	114	54	20	4	132	2	110	2	338	71	47	25	46	10	51	14	137	40	82	7
26-sep	115	61	13	17	14	2	117	3	140	1	370	36	45	12	23	13	149	7	33	31	51	9
27-sep	652	53	51	21	14	3	74	5	295	4	259	54	10	145	23	24	37	58	45	18	19	21
28-sep	568	44	7	10	36	0	65	0	77	0	226	17	3	86	9	24	37	36	79	12	19	3
29-sep	394	1729	28	4	1695	0	62	1	34	0	87	18	9	1715	9	24	25	1690	250	1	692	7
30-sep	566	221	4	3	209	0	60	1	44	1	59	18	237	203	14	6	15	211	39	5	100	3

title	justin	finally	monday motivation	division	entire	kids	today	admitting	difference	y	platform	daily	beats	htt	gone	shirt	reminder	stlouis	kill	theresistance	debunked	nflboycott
15-sep	0	9	3	0	2	18	255	1	6	29	0	12	1	36	515	53	9	192	113	105	2	8
16-sep	0	14	3	4	5	246	109	0	4	11	0	17	0	174	636	34	8	604	195	65	0	15
17-sep	0	5	0	2	0	148	65	0	1	5	0	4	0	33	447	24	8	183	52	34	0	21
18-sep	0	5	16	1	2	31	75	0	3	12	5	8	1	41	62	34	3	268	39	57	1	5
19-sep	0	7	1	0	2	14	153	0	7	13	1	7	0	21	20	441	6	132	44	151	0	5
20-sep	1	9	0	2	4	15	38	0	1571	12	1569	12	1564	77	13	607	10	70	26	144	1	7
21-sep	3	6	0	0	6	43	70	0	195	7	191	6	190	26	13	64	4	50	20	58	0	0
22-sep	1096	5	0	0	2	33	51	0	12	12	16	20	12	4	5	42	4	32	8	45	1	2
23-sep	656	5	0	2	27	456	70	0	4	16	5	9	1	20	17	39	9	31	28	69	0	35
24-sep	136	19	1	5	10	157	221	0	18	30	12	30	9	28	14	33	15	90	21	217	2	989
25-sep	31	8	42	2	93	553	94	0	3	384	8	3	1	25	6	65	18	33	19	108	16	334
26-sep	44	13	96	2	1314	124	80	1	6	349	2	1200	1	30	2	123	1209	34	931	303	1204	79
27-sep	33	7	571	0	384	26	46	0	14	461	1	389	3	45	13	93	391	15	203	177	382	43
28-sep	9	3	497	26	43	21	42	1	3	100	3	42	4	235	3	45	42	13	32	88	29	45
29-sep	4	1686	366	1688	16	5	274	1679	5	297	1	18	0	500	4	39	19	3	18	58	13	33
30-sep	2	204	373	207	9	17	262	198	1	112	1	35	0	490	4	31	10	8	9	48	6	29

title	premise	underpins	maga	u	injured	jasonstockley	players	alllivesmatter	powerful	resistance	court	ajplus	cop	boycottnfl	mad	anthem	w	adamsteinbaugh	displeased	american	f	case
15-sep	0	0	26	45	0	835	1	42	6	60	7	0	133	20	20	8	35	0	0	17	57	44
16-sep	0	0	54	132	51	353	2	88	49	84	4	0	260	9	6	7	36	0	0	53	26	107
17-sep	0	0	59	82	4	135	3	135	7	45	1	0	39	22	5	11	54	0	0	17	18	15
18-sep	0	0	56	53	1	107	1	87	9	45	1	0	24	13	5	13	34	0	0	25	6	9
19-sep	0	0	72	36	6	16	1	68	11	165	1	774	45	19	11	21	118	0	1	43	410	1
20-sep	0	0	363	39	2	34	7	56	6	139	1	71	32	62	3	20	197	0	0	82	568	2
21-sep	0	0	104	50	0	18	3	34	9	130	3	116	47	59	5	16	57	0	0	44	50	5
22-sep	1	0	25	27	0	19	5	40	35	111	22	17	13	43	24	5	20	0	0	15	45	11
23-sep	0	0	74	82	0	3	164	38	6	109	5	5	18	73	8	60	41	0	0	45	27	8
24-sep	2	0	60	98	3	0	238	74	861	116	3	11	79	175	17	605	70	0	0	501	23	8
25-sep	0	0	258	122	1	4	122	541	262	237	2	11	48	178	286	326	140	0	0	366	28	12
26-sep	1195	1194	66	57	0	3	140	146	53	84	0	15	23	118	250	133	180	0	0	79	15	910
27-sep	379	381	73	66	0	1	46	80	20	103	1	3	19	75	360	83	127	0	0	52	19	193
28-sep	28	28	81	324	8	0	45	34	13	32	9	268	300	193	65	57	130	5	5	27	75	33
29-sep	11	11	71	341	1334	0	419	38	17	50	1329	216	346	158	312	70	105	1315	1313	37	54	27
30-sep	6	6	167	54	134	16	341	32	161	14	131	8	75	262	98	33	115	129	129	42	18	42

title	speaks	policebru tality	time	protests	http	ones	vet	ijessewill iams	takeakne enfl	thedudel ukez	smith	truth	acquittal	says	shoot	art	breaking	unarmed	b	Column1
15-sep	2	105	112	39	48	27	1	2	0	0	884	42	891	33	40	3	18	13	43	0
16-sep	12	46	124	112	111	7	0	1	0	0	182	26	237	113	32	25	805	18	36	0
17-sep	28	414	60	92	29	4	1	0	0	0	44	20	81	22	13	5	104	28	19	0
18-sep	7	36	63	67	43	5	0	2	0	0	20	20	33	15	13	12	32	6	11	143
19-sep	10	18	72	193	48	0	513	0	0	0	25	18	15	31	787	7	38	784	66	402
20-sep	6	23	35	140	52	6	123	0	0	0	10	25	8	23	77	4	11	72	101	569
21-sep	3	34	40	57	27	6	31	0	0	0	4	15	4	20	135	6	11	117	43	46
22-sep	487	32	38	68	34	2	8	0	0	0	2	41	2	13	29	8	3	20	8	7
23-sep	339	59	54	67	21	1	14	937	67	0	2	18	2	61	11	9	23	17	37	24
24-sep	81	108	108	100	26	5	77	170	303	0	38	51	3	30	15	854	12	25	166	6
25-sep	36	180	150	159	64	11	491	210	169	0	53	36	0	31	32	243	128	26	104	1
26-sep	101	200	231	105	157	386	76	13	112	0	17	165	0	37	33	43	17	41	41	30
27-sep	144	84	158	95	303	81	11	13	333	104	3	330	0	13	12	20	29	12	92	7
28-sep	36	28	53	45	116	52	4	1	287	497	3	243	0	311	25	8	23	13	41	1
29-sep	15	25	60	38	84	685	4	2	37	366	0	50	0	447	9	10	4	41	29	11
30-sep	118	27	58	25	236	96	1	3	39	373	1	182	0	68	4	6	4	28	423	4

Bigrams

	st_louis	in_st	of_the	trump_supporters	lives_matter	police_officer	this_is	black_lives	in_the	a_police	is_a	police_officers	smash_windows	windows_and	be_sued	he_was	facebook_ads	if_you	does_that	police_cars	and_bricks	rocks_and	i_don't	attack_police	thugs_in
sep	99126	124	44	1123	1	73	243	301	93	139	8	166	38	0	1	0	14	0	876	4	0	0	0	18	2
sep	150569	2722	1677	365	5	130	36	244	150	234	304	274	1467	1643	1642	0	9	0	210	192	1350	1346	1346	37	1348
sep	93272	1188	928	234	1	167	13	232	185	93	10	138	714	683	682	0	4	0	97	143	675	680	680	26	676
sep	73178	652	378	143	2	141	8	158	130	61	9	90	132	119	119	0	13	0	51	17	118	172	172	12	118
sep	75729	483	245	94	8	74	783	98	72	44	780	58	95	64	64	0	466	0	37	6	128	64	64	9	63
sep	83361	204	153	61	3674	110	76	140	136	58	70	54	28	11	11	0	47	0	32	4	17	11	11	7	11
sep	64510	710	503	42	602	66	119	357	64	93	116	48	8	6	6	0	12	0	16	1	7	6	6	10	6
sep	55590	614	530	44	49	75	22	168	76	78	17	26	3	2	2	0	6	0	22	0	3	2	2	10	3
sep	76291	187	169	69	17	50	7	117	62	103	6	99	9	3	3	0	19	0	35	10	3	3	3	10	3
sep	131217	171	74	371	30	509	12	237	576	179	18	226	6	1	1	0	91	0	81	3	1	1	1	65	1
sep	117394	50	44	340	7	239	11	255	275	156	10	242	15	2	2	0	226	234	72	2	2	2	2	27	2
sep	136915	29	73	192	2	158	17	220	93	1023	17	326	64	1	1	0	28	155	115	95	1	1	1	142	1
sep	95161	30	27	436	0	636	7	252	78	311	2	95	13	3	3	0	25	18	172	570	3	3	3	599	3
sep	87992	5	8	405	4	647	14	123	528	105	11	165	18	0	0	667	10	79	44	498	0	0	0	515	0
sep	149210	10	3	114	0	688	1715	150	447	144	1381	310	6	0	0	1237	1323	1723	291	424	0	0	0	372	9
sep	83519	63	53	567	0	457	226	76	108	83	132	378	5	0	0	522	133	203	205	375	0	0	0	394	0

	and_attack	louis_riot	riot_damage	cars_smash	damage_police	with_rocks	officers_with	all_lives	a_racist	black_men	i'm_a	blm_is	that_the	i_guess	don't_support	that_mean	saying_russia	russia_bought	racial_division	takeaknee_mondaymotivation	matter_takeaknee	mean_i'm	racist_all	ads_for	cnn_is
sep	0	0	0	0	0	0	0	0	20	23	15	10	1	24	5	1	3	0	0	0	0	0	0	0	0
sep	1350	1348	1349	1348	1348	1348	1347	1347	27	24	47	13	3	24	10	0	0	0	0	0	0	0	0	0	0
sep	676	675	675	675	675	675	675	675	17	9	59	10	7	2	2	1	0	0	0	0	0	0	0	0	0
sep	125	119	118	118	118	118	118	118	9	8	10	3	1	152	9	1	0	0	0	0	0	0	0	0	0
sep	64	63	63	63	63	63	63	63	11	2	11	2	1	12	5	0	1	0	0	0	0	0	0	0	0
sep	11	11	11	11	11	11	11	11	7	8	37	3	4	10	6	0	1	0	0	0	0	0	0	0	0
sep	6	6	6	6	6	6	6	6	19	7	598	2	2	6	1	1	1	0	0	0	0	0	0	0	0
sep	2	2	2	2	2	2	2	2	7	2	183	4	3	4	3	0	0	0	0	0	0	0	0	0	0
sep	3	3	3	3	3	3	3	3	11	21	713	10	6	4	4	2	0	0	0	0	0	0	0	0	0
sep	3	1	1	1	1	1	1	1	31	11	261	11	1	14	13	7	2	0	0	0	0	1	1	0	0
sep	2	2	2	2	2	2	2	2	23	12	54	5	1	38	5	6	0	0	0	1	1	0	0	0	0
sep	1	1	1	1	1	1	1	1	130	111	25	95	4	1206	6	95	95	0	3	0	93	93	93	93	0
sep	3	3	3	3	3	3	3	3	586	582	46	585	44	389	6	571	570	0	2	0	570	570	570	570	1
sep	0	0	0	0	0	0	0	0	506	507	14	499	21	33	3	498	497	32	10	13	497	497	497	497	10
sep	0	0	0	0	0	0	0	0	374	390	5	369	1675	35	1679	367	381	1692	1696	1685	366	366	366	366	1685
sep	0	0	0	0	0	0	0	0	381	391	13	374	199	19	202	374	373	198	199	205	373	373	373	373	198

all	by_saying	to_cause	bought_facebook	is_finally	cause_racial	is_the	guess_cnn	the_entire	finally_admitting	admitting_blm	has_been	division_i	markdice_by	let_them	a_platform	beats_up	supporters_give	up_trump	amp_let	big_difference	supporters_trump	platform_amp	antifa_beats	them_speak	speak_big
15-sep	0	0	0	0	0	0	54	0	2	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
16-sep	0	0	1	0	1	0	97	0	2	0	0	110	0	0	3	0	0	0	0	0	0	0	0	0	0
17-sep	0	2	0	0	0	0	47	0	0	0	0	36	0	0	0	0	0	0	0	0	1	0	0	0	0
18-sep	0	1	0	0	0	0	45	0	2	0	0	4	0	0	0	3	0	0	0	0	0	0	0	0	0
19-sep	0	0	0	0	0	0	16	0	2	0	0	5	0	0	1	0	0	1	0	0	0	0	0	0	1
20-sep	0	0	1	0	0	0	33	0	0	0	0	5	0	0	1564	1568	1564	1564	1564	1564	1555	1564	1564	1563	1560
21-sep	1	0	0	0	0	0	41	0	2	0	0	0	0	0	190	190	190	190	190	190	190	190	190	190	190
22-sep	0	0	0	0	0	0	96	0	2	0	0	0	0	0	13	12	12	12	12	12	12	12	12	12	12
23-sep	0	0	0	0	0	0	43	0	15	0	0	3	0	0	3	3	1	2	1	2	1	1	1	1	1
24-sep	1	0	0	0	0	0	647	0	6	0	0	19	0	0	9	8	9	6	6	6	9	6	6	6	6
25-sep	0	3	0	0	0	0	346	0	87	0	0	9	0	0	5	0	0	0	1	0	0	0	0	0	0
26-sep	1	0	1	0	1	0	228	0	1310	0	0	1207	0	0	5	1	0	0	0	1	0	0	0	0	0
27-sep	0	0	0	1	0	0	78	0	383	0	0	383	0	0	2	0	0	0	0	0	5	0	0	0	0
28-sep	1	1	0	0	0	0	33	0	43	0	0	38	0	0	3	2	0	0	0	0	0	0	0	0	0
29-sep	1686	1684	1686	1687	1682	1684	43	1682	15	1679	1679	23	1676	1675	1	0	0	0	1	0	1	0	0	0	0
30-sep	201	198	198	198	198	198	34	198	7	198	198	20	198	198	1	0	0	0	0	0	0	0	0	0	0

all	prisonplanet_antifa	reminder_that	takeaknee_and	of_black	you_are	to_be	daily_reminder	your_daily	underpins_takeaknee	entire_premise	which_underpins	been_debunked	premise_which	prisonplanet_your	sued_deray	protest_the	for_the	can't_be	a_protest	because_he	officer_who	during_a	the_court	deray_and/n	Column1
15-sep	0	0	6	0	49	861	40	2	2	0	0	0	0	0	0	0	2	97	1	0	2	838	3	0	0/n
16-sep	0	0	2	0	228	190	96	0	0	0	0	0	0	0	0	0	2	144	2	5	2	161	4	2	0/n
17-sep	0	1	3	0	166	55	77	1	1	0	0	0	0	0	0	0	3	48	0	4	1	39	2	0	0/n
18-sep	0	0	1	0	29	41	89	1	1	0	0	0	0	0	0	0	1	32	8	2	10	21	3	1	4/n
19-sep	0	0	5	0	9	39	40	1	1	0	0	0	0	0	0	0	0	51	4	0	3	14	3	0	3/n
20-sep	1560	1558	5	0	30	124	50	3	1	0	0	0	0	0	0	0	1	35	0	0	0	7	2	0	0/n
21-sep	190	190	1	0	37	24	40	0	0	0	0	0	0	0	0	0	1	47	3	1	6	2	2	1	0/n
22-sep	12	12	3	0	10	15	85	3	3	0	0	0	0	0	0	0	0	32	1	0	2	2	4	20	0/n
23-sep	1	1	5	12	414	50	438	2	2	0	0	0	0	0	0	0	6	52	0	3	3	1	2	3	0/n
24-sep	6	7	4	15	210	64	304	2	2	0	0	0	0	0	0	0	18	82	2	15	0	1	3	0	0/n
25-sep	0	0	11	14	54	68	90	1	1	0	0	0	0	0	0	0	12	91	2	10	8	0	1	0	1/n
26-sep	0	0	1200	1206	264	60	60	1194	1194	1194	1194	1194	1191	1194	1187	0	22	126	1	5	2	1	2	0	1/n
27-sep	0	0	385	383	79	28	49	380	380	381	378	378	381	378	377	0	7	87	2	4	1	2	0	0	3/n
28-sep	0	0	34	33	39	9	46	29	28	28	28	28	28	28	28	63	9	168	598	11	9	103	8	6	6/n
29-sep	0	0	17	12	18	17	78	11	11	11	11	11	11	11	11	1342	1315	96	775	1319	1318	248	1314	1314	1316/n
30-sep	0	0	7	6	24	11	57	6	6	6	6	6	6	6	6	131	133	340	125	130	130	49	132	129	129/n