Introduction to Tools and Methods for the Analysis of Twitter Data

(Journal Article)

Manuel Burghardt

The microblogging service Twitter provides vast amounts of user-generated language data. In this article I give an overview of related work on Twitter as an object of study. I also describe the anatomy of a Twitter message and discuss typical uses of the Twitter platform. The Twitter Application Programming Interface (API) will be introduced in a generic, non-technical way to provide a basic understanding of existing opportunities but also limitations when working with Twitter data. I propose a basic classification system for existing tools that can be used for collecting and analyzing Twitter data and introduce some exemplary tools for each category. Then, I present a more comprehensive workflow for conducting studies with Twitter data, which comprises the following steps: crawling, annotation, analysis and visualization. Finally, I illustrate the generic workflow by describing an exemplary study from the context of social TV research. At the end of the article, the main issues concerning tools and methods for the analysis of Twitter data are briefly addressed.

10PLUS1
LIVING
LINGUISTICS

1. Introduction

New Media and Quantitative Methods in the Digital Humanities

When Lev Manovich speaks of the "language of new media", he does not actually mean language in a linguistic sense, but rather uses the term as an umbrella for different elements that influence new media and thus constitute a language of their own (Manovich 2001: 7). One main characteristic of all new media is its transcoding, i.e. it is represented as computer data and therefore comprises not only a cultural layer, but also a computer layer (45ff.). Due to this computer layer of new media, Burger and Luginbühl (2014: 445) propose "digital media" as an alternative term, including media types such as digital television, smartphones, and the Internet with its various applications and services. The computer layer also allows for new ways of computer-based, quantitative analysis that goes beyond traditional, hermeneutic approaches typically known in the humanities. Accordingly, the term digital humanities is oftentimes used to subsume all kinds of computer-related, empirical methods that can be used in the humanities, including the analysis of new media which is heavily influenced by computers. While corpus linguistics already have a strong tradition of using empirical methods, recent approaches such as "culturomics" (Michel et al. 2011), "distant reading" (Moretti 2007, 2013), and "macroanalysis" (Jockers 2013) are currently being discussed by the literary and cultural studies community as well.

Linguistics and Social Media Language Data

Modern corpus linguistics has been on the rise since the advent of technological innovations such as desktop publishing and the Internet, which essentially resulted in an increased availability of digital, machine-readable language data. While the web as a corpus (cf. Baroni et al. 2009; Kilgarriff & Grefenstette 2009) may be seen as a well-established subfield of corpus linguistics by now, the growing landscape of social media platforms add many new perspectives (as well as challenges) to the field of linguistics. One of the most striking features of social media language data is that it is user-generated, i.e. regular people are communicating with each other. The communication is obviously influenced by the channel, i.e.

communication via the Internet has developed its very own characteristics (cf. Crystal 2006; Beißwenger & Storrer 2008; Marx & Weidacher 2014) that result in specific phenomena in all areas of linguistics, including orthography and lexis as well as syntax, semantics and pragmatics.

Twitter as an Object of Study

In this article I will provide an introduction to the social media platform Twitter and give an overview of tools and methods that can be used to study Twitter data from a media linguistics perspective. Although there are numerous other social media services (e.g. Facebook, Reddit, YouTube, Flickr, etc.), Twitter has quickly become one of the most popular objects of study in the academic community. I believe this is due to a number of characteristics of the Twitter platform:

- **Message size:** Twitter messages are relatively short (comparable to an SMS), which results in relatively homogeneous corpora. In comparison, Facebook posts, emails, or blog posts may vary in length considerably, which makes it more difficult to create balanced, comparable corpora.
- **Sample size:** Several million messages are published on Twitter every day, i.e. it is possible to get large amounts of data, even for very recent events.
- **Metadata:** Twitter messages provide all kinds of metadata, e.g. username, date of creation, language, geolocation, and many more.
- **Availability:** Most Twitter data is publicly available, even for passive users of Twitter, i.e. for people who have no registered Twitter account.
- **Accessibility:** Twitter data can be accessed and downloaded relatively easy via a pre-defined Application Programming Interfaces (API).[1]

An overview of the evolution of Twitter as an object of study is given by Rogers (2013). Williams et al. (2013) provide a classification of academic papers that are dedicated to Twitter, trying to answer the question “What do people study when they study Twitter?”. Twitter research covers a wide range of disciplines that include information behavior (Meier & Elsweiler 2014), sentiment analysis (Pak & Paroubek 2010), and linguistics: A number of articles from the field of computational linguistics and natural language processing deal with the question how part-of-speech tagging for Twitter data can be improved (Gimpel et al. 2011; Derczynski et al. 2013; Rehbein 2013). Zanzotto et al. (2011) analyze linguistic redundancy in the language used on Twitter. González-Ibáñez et al. (2011) describe a corpus of sarcastic Twitter messages and discuss problems for the automatic identification of sarcasm on the lexical and pragmatic level. Han & Baldwin (2011) analyze out-of-vocabulary words that are used on Twitter and suggest an automatic approach for the lexical normalization of such noisy language data. One of the few examples for linguistic research on Twitter that is not corpus-based, but rather provides insights on the conceptual level, can be found in Overbeck (2014), who attempts a text linguistic classification of Twitter data.

[1] Obtaining Twitter data via this API, however, requires some basic programming expertise which may be a hurdle for many scholars without a computing background. At the same time, the Twitter API is rather complex, and brings along some limitations, i.e. it is not possible to download any set of Tweets for an arbitrary time span. One main goal of this article is to give an overview of the basic characteristics of the Twitter API, and to introduce existing tools that can be used to obtain and analyze Tweets without having to do any programming at all.

This article aims to enable even more research on Twitter data by introducing the technical foundations as well as some available tools that allow humanities scholars to gather and analyze Twitter messages.

2. How Twitter Works

Since it was founded in 2006, Twitter has quickly become one of the most popular services in the social media landscape.[2] In June 2015, the company stated they had 302 million active users monthly, who write approximately 500 million posts per day. 80% of Twitter users use the service from a mobile device. Twitter supports over 35 languages. 77% of the registered accounts are outside the U.S., i.e. Twitter posts are available in many different languages[3] (cf. Figure 1).

As a microblogging platform, Twitter incorporates many characteristics that are also known from more traditional blogging software, e.g. the possibility to keep track of other people's blogs or to comment on their posts. A distinctive feature of Twitter when compared to other blogging services is the limited amount of characters available for a message text, which will henceforth be referred to as a Tweet. A Tweet consists of a maximum of 140 characters, and will be described in more detail in the next section. Every registered Twitter user has a personal timeline, which displays their own Tweets as well as Tweets by people they have chosen to follow; these are displayed in chronological order (also cf. Russell 2013: 9ff). It is also possible to display the other users' timeline.

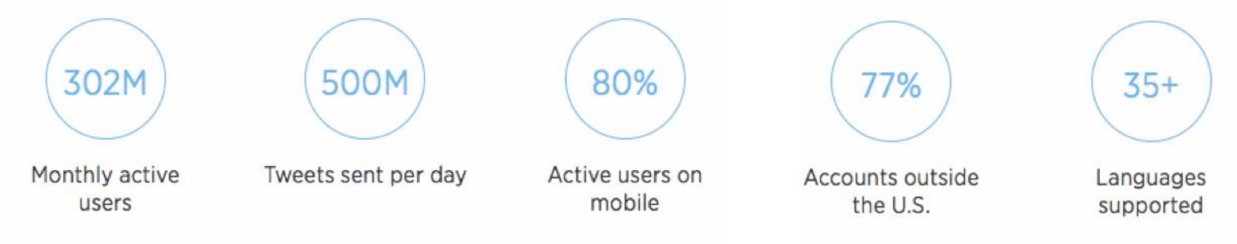


Figure 1: Extract of a infographic provided by Twitter (source: https://about.twitter.com/company, June 1, 2015).

Anatomy of a Tweet

Each Tweet published via Twitter shares the same basic structure and comes with different types of information – some optional, some obligatory. Figure 2 shows a schematic overview of the basic structure of a Tweet.

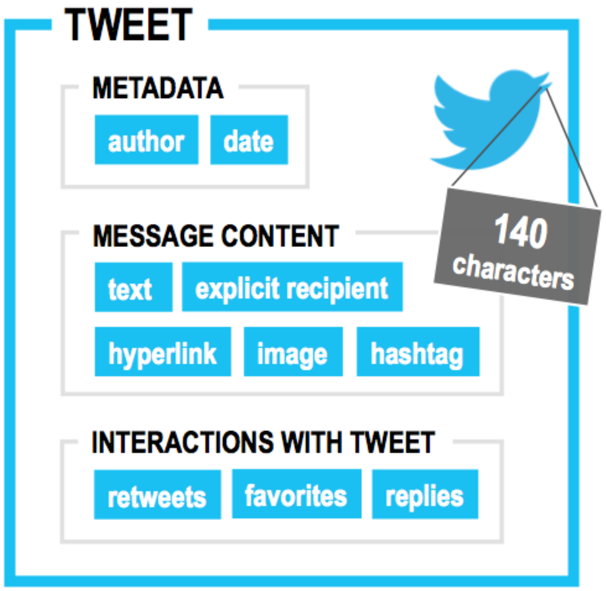


Figure 2:
Schematic overview of the basic structure of a Tweet.

(1) Metadata: A Tweet can only be published via a registered Twitter account, i.e. each Tweet has an explicit author, which is the username. In addition, the exact time and date of publication for Tweets are available.

[2] For a comprehensive overview of the history of Twitter cf. Makice (2009: 9ff).

[3] This information was taken from the official fact sheet of the Twitter company, available at https://about.twitter.com/company; all URLs mentioned in this article were last checked on June 1, 2015.

(2) Message content: The actual message body of a Tweet may contain different types of information. Users may post plain text, hyperlinks, images, or videos. By default, a Tweet is visible for anybody who follows the author of the Tweet in their personal timeline. Users may also choose to send Tweets exclusively to a specific recipient, which is called a "mention" in Twitter.[4] This can be achieved by writing the Twitter username of the recipient in the message text and by putting an @ in front of it (example: "@8urghardt – How are you?"). One of the most distinctive features of Twitter is the use of hashtags. A hashtag can be created by putting a hash (#) in front of any string of characters. The basic idea of hashtags is to provide keywords for a Tweet that describe its basic topic. Twitter can be searched for hashtags, i.e. if you search for "#obama" you will get a list of all Tweets that have been labeled with the corresponding hashtag. It is also possible to provide multiple hashtags for one Tweet. However, not all hashtags are used as descriptors of the topic of a Tweet. In fact, many authors use hashtags to indicate sarcasm or irony, or to express additional conversational information. For a basic introduction into the use and function of hashtags on Twitter see Kricfalusi (2015) and Cunha et al. (2011) for an overview of the "dynamic evolution of hashtags on Twitter".

(3) Interactions with Tweet: Once a Tweet has been published into the Twittersphere, other users have several ways to interact with the Tweet. Retweeting a Tweet means posting a Tweet that has already been published (typically by somebody else) to one's own list of followers. Retweeting is usually considered a means of showing appreciation for a Tweet, as it makes it accessible for a wider circle of people. Retweeting is a basic mechanism for viral network effects. Users may also add a Tweet to their list of favorites, which resembles the bookmarking mechanism of web browsers. What exactly users are trying to achieve or to communicate when they retweet (boyd et al. 2010) or favorite (Meier et al. 2014) a Tweet has been

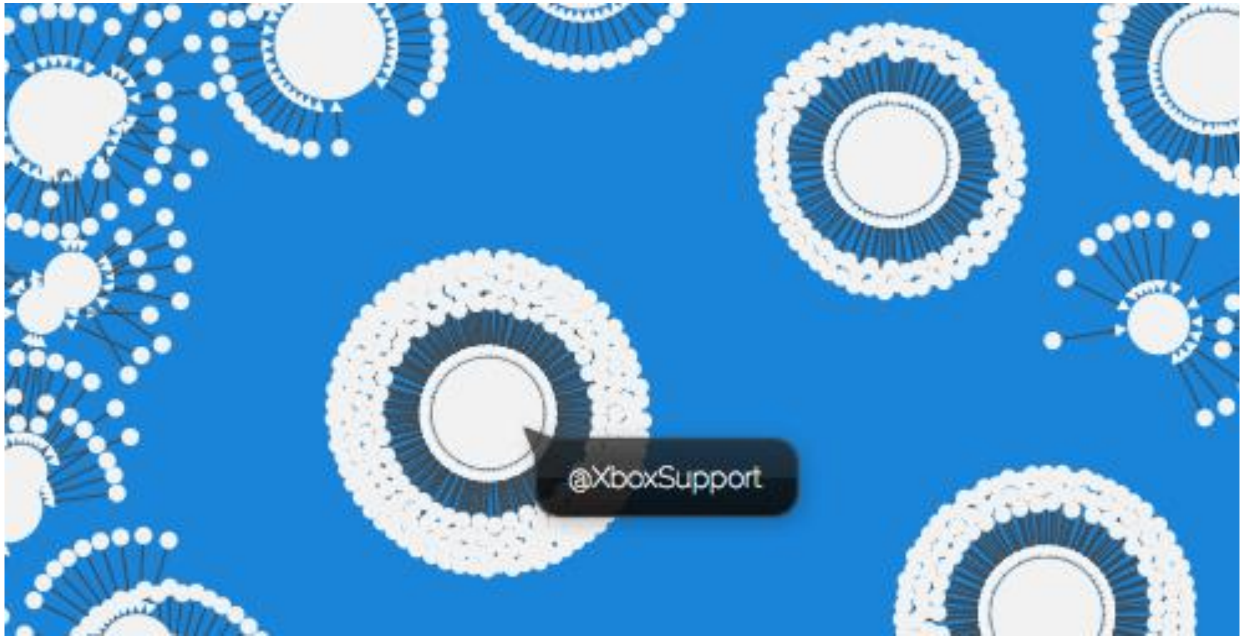


Figure 3: Reply network for Tweets that contain the hashtag #xbox.
Image taken from the Twista analysis and visualization tool for Tweets (Spanner et al. 2015).

[4] For more details about mentions in Twitter cf. https://media.twitter.com/best-practice/what-are-replies-and-mentions.

a continuing research topic in the area of information behavior and personal information management. Finally, users may reply to a Tweet with a Tweet of their own, thus creating a dialog between two or more Twitter users. Figure 3 shows a visualization of a reply network for different Twitter users who have posted Tweets about the keyword “xbox”. It becomes obvious that a few users, such as the official XboxSupport account, are the central communicators in the network (depicted as larger bubbles), and that many other users (depicted as smaller bubbles) reply to messages from these central users.

Twitter Usage

Twitter is used by pop stars and actors, by customers and companies (Jansen et al. 2009), by academics (Ross et al. 2011) and politicians (Ausserhofer & Maireder 2012). A common goal of publishing a Tweet is not only to share ideas and opinions, but also digital resources in the form of web links, images and videos. Java et al. (2007) analyzed the usage of Twitter systematically, and came to the conclusion that the main functions of Tweets are daily chatter, conversations, sharing information / URLs, and reporting news.

3. Collecting Twitter Data: The Twitter API and Available Tools

Limitations for Collecting Tweets

It is important to note that according to Twitter’s terms of use,[5] redistributing Twitter content outside the Twitter platform is prohibited. In practice, this means it is not possible to precompile Tweet corpora and to share them in a way they are readily accessible for academic research. A workaround for these limitations that can be used to share corpora of Tweets with others nevertheless is described by McCreadie et al. (2012): Tweet corpora may be shared as a list of numerical identifiers (IDs) that can be used to reconstruct Tweet content via the Twitter API. The Twitter API is a pre-defined interface with which developers can communicate with the Twitter platform. This approach is, however, rather impractical, as it involves basic programming skills to build Tweets within the API by using their IDs as input. Another problem here is that Tweets that are reconstructed via their ID may change through the course of time, i.e. they may be deleted, their message content may be modified, and, of course, the number of retweets and favorites may change. For an example of this type of available Tweet ID corpora, cf. the TREC 2011 Microblog Dataset.[6]

This essentially means that there are no readily available corpora of Tweets that can immediately be used for academic studies. Rather, scholars are required to create their own collections of Tweets via the Twitter API. However, there are a number of tools and services that provide a graphical user interface for the Twitter API. In the remainder of this chapter, I will quickly introduce the Twitter API to illustrate what kind of information can be obtained, but also which limitations exist for collecting Twitter data. In the last part, some available tools that can be used to create tailored Tweet corpora will be introduced.

[5] Twitter Developer Agreement: https://dev.twitter.com/overview/terms/agreement-and-policy.

[6] TREC Tweets2011: http://trec.nist.gov/data/tweets/.

The Twitter API

Via the Twitter API it is possible to query a number of different parameters for the basic objects user, Tweet, entity and place. An overview of the most important types of information that are available via the API is displayed in Table 1. From a linguistic perspective, the object Tweet is most relevant, as it not only contains the message text but also a number of other relevant parameters such as language and geolocation. The language used in a Tweet is assessed on Twitter by a language-detection algorithm. It is important to note that any kind of geo-information is only available when the author of a Tweet has geo-tagging enabled, i.e. this information will not be available for all Tweets.

In order to be able to request Tweet data via the Twitter API, the user has to authenticate himself via the Open Authentication (OAuth) mechanism, which in turn requires the registration of an application on the Twitter platform[7] beforehand. A comprehensive overview of the authentication process is given in Kumar et al. (2013: 6-7).

Object	Types of information
User	Username, screen name, user description text, favorites count, followers count, friends count, user-declared language
Tweet	Message text, number of retweets, number of favorites, date of creation, machine-detected language, geo-coordinates
Entity	Hashtags, uploaded media, URLs, user mentions
Place	Country name, city name, type of location

Table 1:
Overview of basic types of information available via the Twitter API (taken from the Twitter API Overview, available at https://dev.twitter.com/overview/api).

Twitter essentially provides two different types of APIs, which can be used to achieve rather different things.

(1) The Search API,[8] which is part of Twitter's REST API, can be used to explicitly search for Tweets that match a specified criterion (e.g. a keyword, hashtag, or username), and behaves similarly like the Twitter.com online search function.[9] It is important to note that the Search API does not provide access to all past Tweets, but only includes Tweets from the last 6-9 days.

(2) The Streaming API can be used to get access to a continuous stream of newly published Tweets. These Tweets can be filtered by different parameters such as keywords, geolocation or user ID. The Streaming API returns all Tweets that match those filter criteria up to a volume that does not exceed 1% of the total current volume of Tweets published on Twitter (Kumar et al. 2013: 20).

It is important to be aware of these basic limitations of the different API types when using them to obtain Twitter data (cf. Figure 4).

Existing Tools for Collecting and Analyzing Twitter Data

The landscape of software tools that can be used for the analysis of Twitter is vast and diverse. A basic way to categorize tools is by means of their analytic focus: A great number of Twitter tools are dedicated to social media analytics, i.e. they focus on social networks of Twitter users (e.g. follower growth) and how successful a Tweet is distributed in

[7] Create a new Twitter application: https://apps.twitter.com.

[8] All information about the Search API in this paragraph was gathered from the official Twitter documentation, available at https://dev.twitter.com/rest/pub-lic/search.

[9] Twitter search function: https://twitter.com/search-home.

the Twittersphere (user-centric tools). Important parameters for these analyses are follower counts, retweet counts and favorite counts. Twitonomy,[10] Twittercounter,[11] MyTopTweet,[12] Riffle,[13] and TweetReach[14] are among these tools, but there are also more generic social media analytics tools such as Sumall,[15] which not only allow users to monitor Twitter, but also other services such as Facebook, Instagram, YouTube, and many more. The other class of tools is more focused on obtaining and analyzing the message text and available metadata of the actual Tweets (Tweet-centric tools). In this article I will primarily focus on Tweet-centric tools, as they are more suited for transferring corpus linguistic methods to Twitter data than user-centric tools.

While it is possible to create a custom computer program that makes use of one of the Twitter APIs to obtain Tweets, there is also a great number of tools that are readily available and that can be used to obtain tweets via the Twitter API. Accordingly, Twitter tools can be further distinguished by the specific type of API they are utilizing.

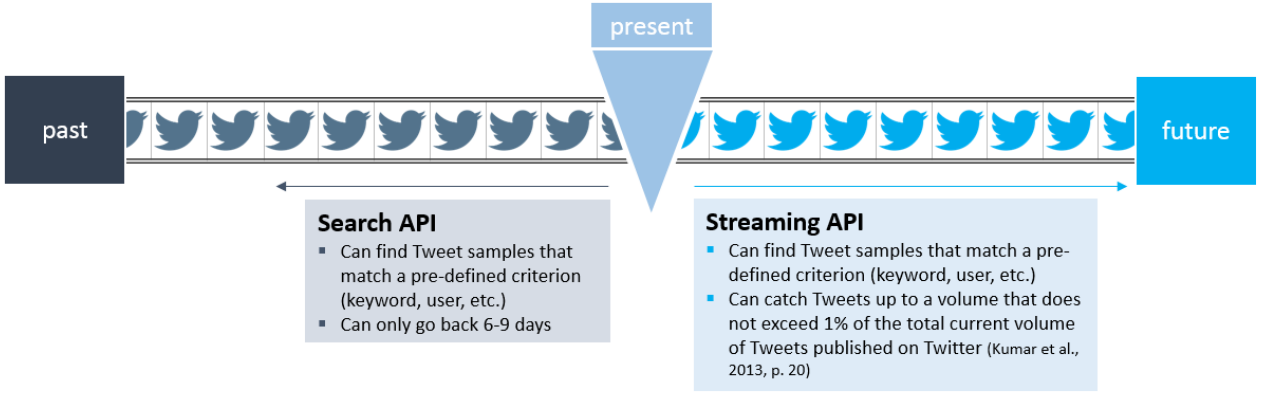

present
past
future
Search API
▪ Can find Tweet samples that match a pre-defined criterion (keyword, user, etc.)
▪ Can only go back 6-9 days
Streaming API
▪ Can find Tweet samples that match a pre-defined criterion (keyword, user, etc.)
▪ Can catch Tweets up to a volume that does not exceed 1% of the total current volume of Tweets published on Twitter (Kumar et al., 2013, p. 20)

Figure 4: Illustration of key differences between the Twitter APIs.

(1) Firehose tools – There are a few services, such as Gnip[16] and Topsy,[17] that have the status of a certified reseller of Twitter data, i.e. these companies pay Twitter to get access to *all* Tweets that have ever been published via a specific variant of the Streaming API which is called Firehose.[18] As it is the business model of these companies to provide a searchable structure for billions of tweets, they are not free of charge, but are rather intended for the commercial business analytics sector. To get an idea of what kind of data such services can provide, Topsy can be tried out for free, but only returns Tweets from the past 30 days, and only displays the top 100 tweets of a search (Wagner 2013). It can also be used to visualize the diachronic development of one or more concepts with regard to the number of Tweets that mention a specific concept (cf. Figure 5).

(2) Streaming API tools – As aforementioned, the Streaming API allows developers to tap the continuous stream of newly published Tweets and to store those Tweets up to an extent of 1% of the overall Twitter traffic (approx. 500 million Tweets per day). Twista (Spanner et al. 2015) is an example of a tool that uses the Streaming API to collect Tweets that match pre-defined criteria, e.g. hashtags or keywords, for a specified period

[10] http://www.twitonomy.com.
[11] http://twittercounter.com.
[12] https://mytoptweet.com.
[13] http://crowdriff.com/riffle.
[14] https://tweetreach.com.
[15] https://sumall.com.
[16] https://gnip.com.
[17] http://topsy.com.
[18] https://dev.twitter.com/streaming/firehose.

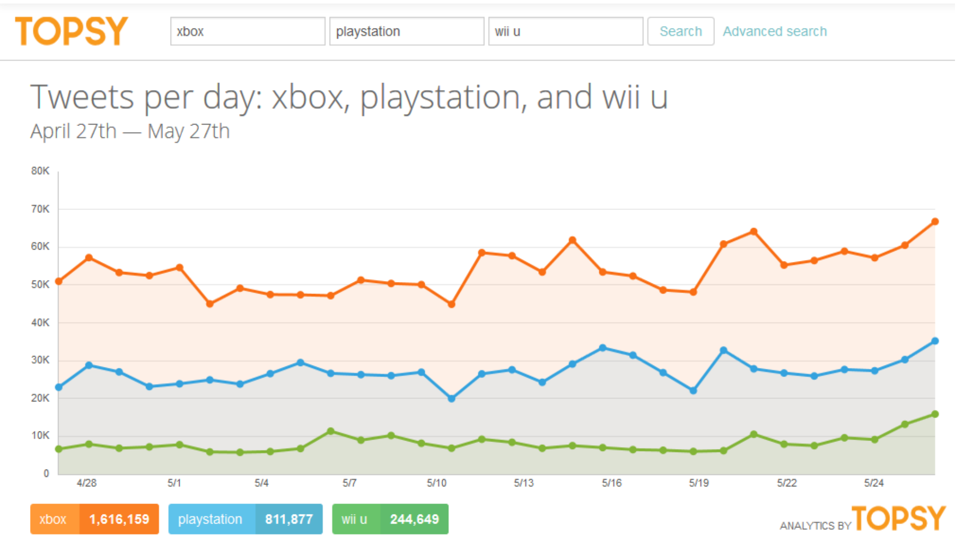


Figure 5: Frequency distributions for the concepts xbox, playstation and wii u in all public Tweets over the course of one month (cf. http://topsy.com/).

of time. Once the specified crawling period is over, the user gets notified via email that their collection is ready for download. On top of the crawling, Twista also provides a number of content analytics and visualizations for the collected Tweet corpus which may be displayed interactively in the web browser (cf. Figure 6).[19]

Another existing tool that makes use of the Streaming API is Tworpus[20] (Bazo et al. 2013) . Tworpus is a service that continually collects as many Tweets as possible and stores them in an internal database. Started sometime in 2013, more than 300 million Tweets in 8 different languages have been collected thus far (cf. Figure 7).

Users can create tailored corpora by specifying the following parameters:

- corpus size, i.e. total number of Tweets
- language(s) used in the Tweets
- period of time for the Tweet publication date
- minimum / maximum number of characters used in Tweets

The corpus is then built from the Tweets stored in the database and can be downloaded for further analyses in XML or plain text format. As previously mentioned, Twitter does not allow developers to redistribute Tweets outside of the Twitter platform. Tworpus therefore makes use of the concept described by McCreadie et al. (2012), i.e. not the Tweets themselves are stored, but rather their unique identifiers and corresponding metadata. The actual Tweet corpus is then built by resolving the IDs and fetching the actual Tweets from Twitter (Bazo et al. 2013).

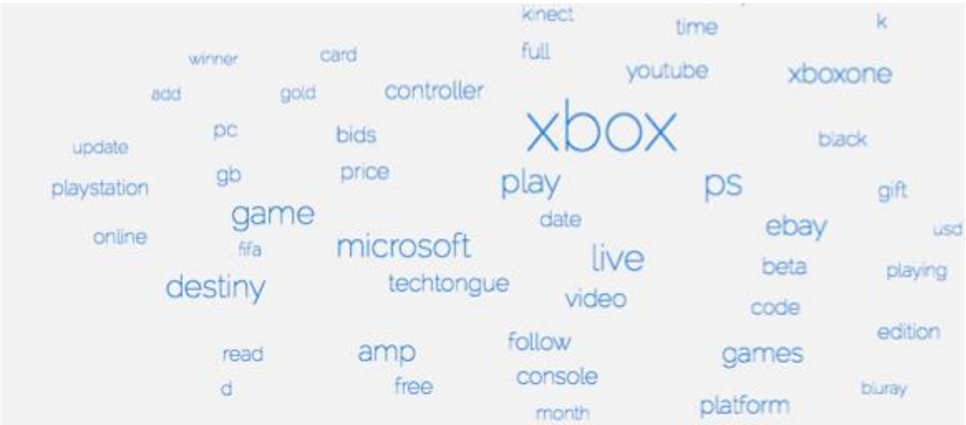


Figure 6: Visualization of the most frequently used words for a collection of tweets that contain the keyword xbox (example taken from Spanner et al. 2015).

[19] Right now, Twista has the status of a working prototype that is not yet publicly available. Parties interested in using the tool should contact the developers directly (cf. Spanner et al. 2015). An example corpus with corresponding analyses is available at http://bit.ly/1xephsf.

[20] http://tools.mi.ur.de/tworpus.

(3) Search API tools – Another category of tools allows users to download collections of Tweets by means of the Search API, which means that Tweets can only be looked up in the past if they are not older than 6-9 days. Among these tools are commercial variants such as TweetArchivist,[21] but also free-of-charge tools such as Martin Hawksey's TAGS (Twitter Archiving Google Sheet).[22] These tools can also be used to monitor Tweets for a certain keyword or hashtag by automatically querying the Search API in user-defined intervals (e.g. every hour). Tweet collections from such tools can typically be downloaded as CSV (comma separated values) file or as an Excel spreadsheet that can then be used for further analyses. TAGS also provides an explorer component that can be used to analyze and visualize the collected data right away (cf. Figure 8).

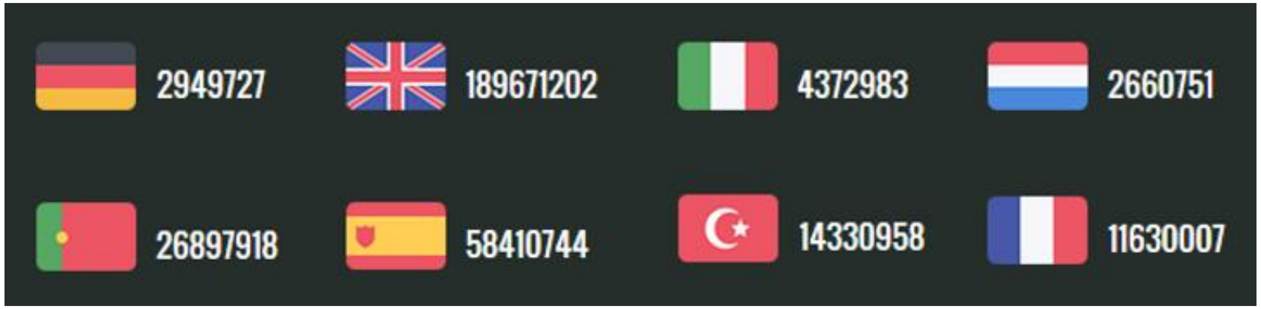


Figure 7: Overview of Tweets Crawled by Tworpus (June 1, 2015).

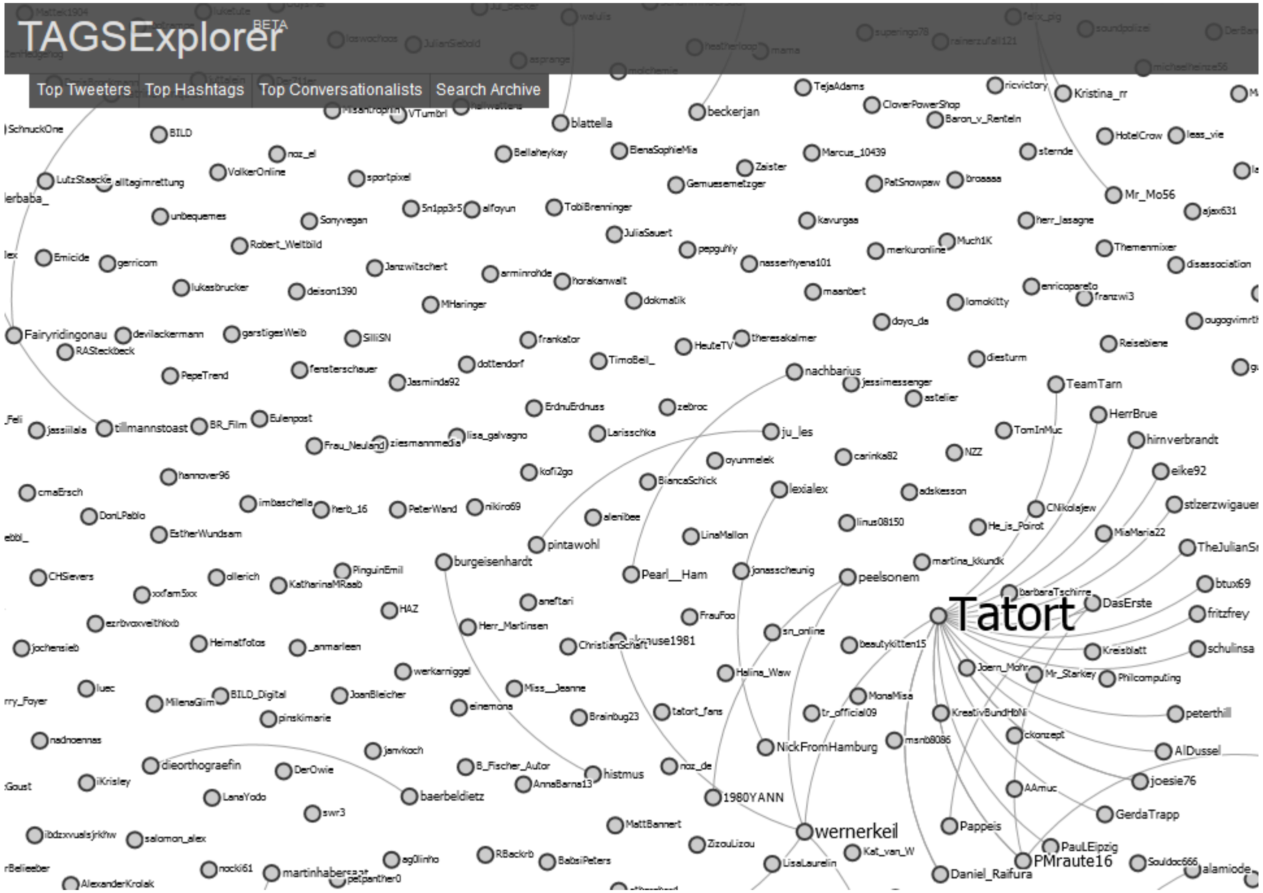


Figure 8: TAGSExplorer visualization of the network of users that have published Tweets with the hashtag #tatort.

[21] https://www.tweetarchivist.com.
[22] https://tags.hawksey.info.

4. A Basic Workflow for Conducting Studies with Twitter Data

So far, the ways in which Tweets can be obtained from the Twitter API by using currently existing tools has been demonstrated. In this chapter I will present a more comprehensive workflow that includes all steps that are necessary to conduct a study with Twitter data, and that also suggests existing tools for the realization of each step (cf. Figure 9).

(1) Crawling – While obtaining actual Tweet data, which is oftentimes called crawling, is logically the first step, it is only the beginning of a more complex research workflow that includes analyzing and interpreting the data.

(2) Annotation – Although Tweets already come with a number of interesting metadata (cf. Table 1), it may be helpful to add further annotations, e.g. the gender of the author of a Tweet, or a descriptive content category (e.g. conversational Tweet,

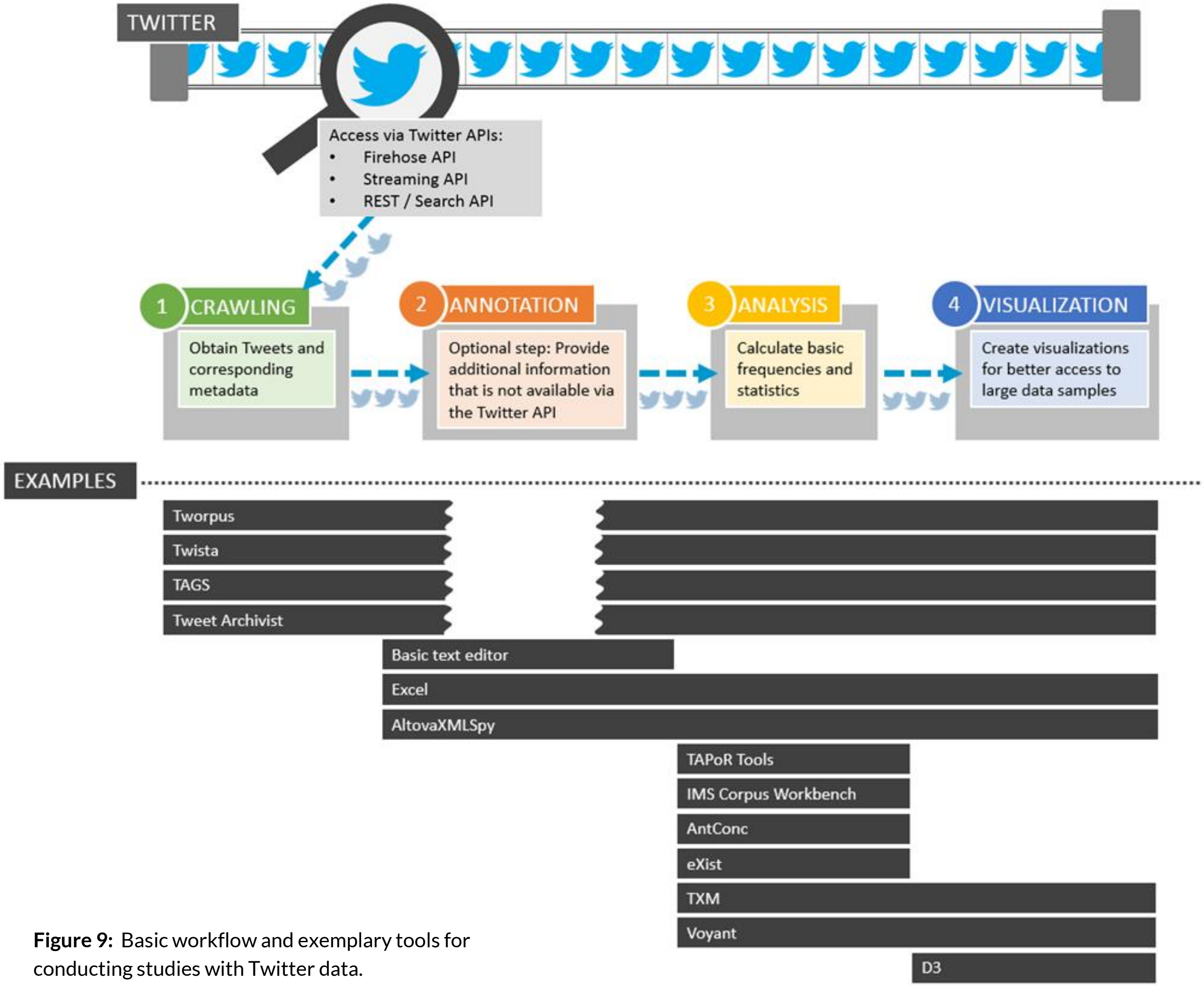


Figure 9: Basic workflow and exemplary tools for conducting studies with Twitter data.

ironic Tweet, etc.). Additional annotations are an optional step in the workflow. Technically, annotations are added to the data structure of the collected Tweet corpus, which is usually in JSON (JavaScript Object Notation), XML (eXtensible Markup Language), or CSV (Comma Separated Values) format. A helpful tool for the transformation of data from one format into another is the online tool DataWrangler.[23]

(3) Analysis – Research in the context of Twitter often uses large sample sizes, as Tweets are easily accessible and can be obtained in large numbers; therefore, semi-automatic, quantitative analyses typically accompany Twitter studies. Many crawling tools already provide basic analysis components that count the most frequent authors, frequently used words and hashtags, and other basic frequencies. There are, however, also a number of dedicated tools that can be used to perform quantitative analyses of Tweet corpora as well. For corpora in XML format, existing analysis tools are the TAPoR Tools,[24] IMS Corpus Workbench,[25] TXM[26] or eXist.[27] For data in plain text format, tools such as AntConc[28] or VoyantTools[29] can be used to analyze the data.

(4) Visualization – Adequate information visualization is an important requirement for Twitter analysis tools, as visual representations have specific advantages in comparison to written text when it comes to making large data sets more accessible (Larkin & Simon 1987; Mazza 2009). Some of the aforementioned crawling tools also provide basic visualizations. Examples for tools that can be used to visualize Tweet data outside of existing crawling tools include VoyantTools, which provides a number of different visualization options,[30] or the D3 (Data Driven Documents)[31] framework, which is ideal for the visualization of JSON data, but requires knowledge of HTML and JavaScript. Altova XMLSpy[32] is an exemplary tool that can be used to analyze XML data and to create diagrams and other types of visualizations. For an overview of possible visualization techniques for different aspects of Tweet data, which include the visualization of network information, temporal information, geo-spatial information and textual information (cf. Kumar et al. 2013: ch. 5).

5. Example Study: Social TV – Twitter and the *Tatort* Series

In this chapter I present an example study conducted in the context of social TV and Twitter usage (Burghardt et al. 2013). I will illustrate how the basic workflow from the previous chapter can be implemented in an actual study and will also provide a quick hands-on guide on the particular tools used in this study.

Twitter and the Tatort Series

Proulx and Shepatin (2012: 11) observe that social media services such as Facebook and Twitter are being used in the context of a social TV experience more often, as they can be used to provide an interactive backchannel for the traditionally rather static TV scenario (see also Klemm & Michel in the current issue). Accordingly, Twitter usage during one of the most popular detective series

[23] http://vis.stanford.edu/wrangler.
[24] http://taporware.ualberta.ca.
[25] http://cwb.sourceforge.net/index.php.
[26] http://sourceforge.net/projects/txm.
[27] http://exist-db.org/exist/apps/homepage/index.html.
[28] http://www.laurenceanthony.net/software.html.
[29] http://voyant-tools.org.
[30] http://docs.voyant-tools.org/tools.
[31] http://d3js.org.
[32] http://www.altova.com/de/xmlspy.html.

in the German TV landscape, *Tatort*, was analyzed. *Tatort* has been around since 1970 and is aired every Sunday evening from 8.15-9.45 p.m. Although it is a rather traditional TV series, there is an active community of people who publish live Tweets about the show while it is being broadcast. *Tatort* can be seen as a typical case of Twitter being used as an interactive backchannel to create a social TV experience and to communicate with others who watch the program. We have created a corpus of Tweets for one specific episode of *Tatort*, which allows us to analyze the typical functions and contents of *Tatort* Tweets.

Crawling

Tweets about *Tatort* can be easily identified, as they are tagged with the characteristic hashtag #tatort, i.e. any Twitter user who wants to make sure that their Tweet is recognized by the *Tatort* social TV community will use this hashtag. At the same time, the hashtag is rather unique, i.e. it is only rarely used in Tweets that are not connected to the TV series. As the study was focused on Tweets published during the live *Tatort* broadcast, only data in the time frame from 8.15–9.45 p.m. was collected.

Originally, we used TweetArchivist – which was then a freely available tool – to create our corpus. It would have been possible, however, to use the freeware tool TAGS to create the same collection of tweets, as both tools utilize the Search API and allow users to download the corpus in CSV format, which can then be imported into and modified in by spreadsheet programs such as Microsoft Excel. Before TAGS can be used, a new application has to be registered at Twitter.com, i.e. you will need a valid Twitter account. New Twitter applications can be registered at:

https://apps.twitter.com/app/new

In the application details form you can provide an arbitrary name, description and (fictitious) website. It is, however, important to provide the following value for the field “callback URL”:

https://script.google.com/macros/

After receiving the above information, Twitter will generate a Consumer Key (API Key) and a Consumer Secret (API Secret) that are necessary in order to connect the TAGS tool to the newly filed Twitter application. TAGS does not require the installation of any software on a local computer, but rather is an extension to Google’s spreadsheet tools, which are freely available online (but require a valid Google account):

https://accounts.google.com/SignUp

A personal TAGS spreadsheet (TAGS version 6.0) can be created at:

https://tags.hawksey.info/get-tags/

Finally, the last step is to enter the previously generated Consumer Key and the Consumer Secret into the TAGS settings (“TAGS > Setup Twitter Access”). All the steps described in this paragraph are only needed when setting up TAGS for the very first time.

Once TAGS is fully set up, you can enter one or more terms that will be used to filter Tweets (cf. Figure 10). For this case study, we provided “#tatort” as a filter term. The time frame for this study could also have been set in TAGS.

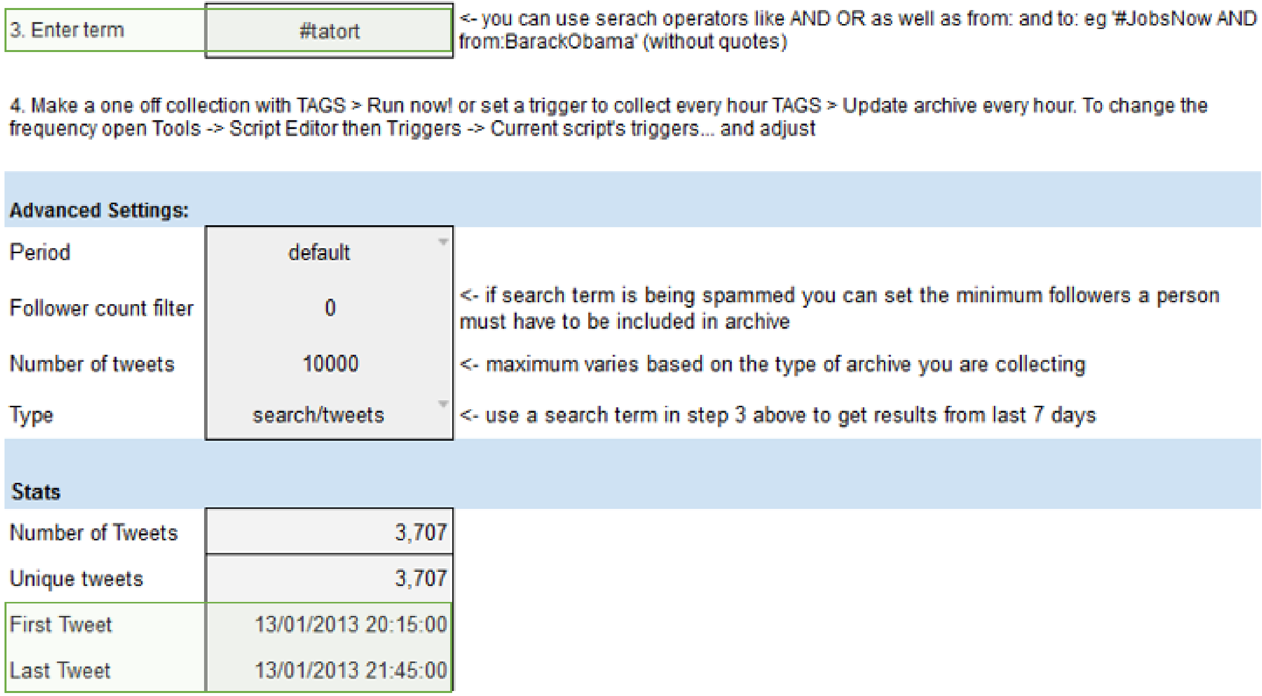
We analyzed the *Tatort* episode 859 (“Kaltblütig”, D 2013, Andreas Senn) and collected a total of 3,707 Tweets for this episode. The results of the query were returned to a Google spreadsheet titled

"Archive" and could be downloaded as a CSV or Excel file for further annotations and analyses. The spreadsheet contains different kinds of information that is available via the Twitter Search API, including the author of the Tweet, the actual message text, hashtags, and much more.

Annotation

In the second step of the workflow, we manually categorized all Tweets by their function. We created the coding scheme by means of a content analysis approach with two independent coders and 100 Tweets from a previous *Tatort* episode. The resulting 14 categories were evaluated by several test persons, which led to a rephrasing of some categories for better comprehensibility. Some categories were evaluated as being too generic, and so we further differentiated them and came to a final set of 17 categories. Some examples for these categories are:

- critique → personal evaluation of the episode
- speculation → speculation about how the plot would further develop
- joke → jokes and puns about the plot or about dialogs



3. Enter term	#tatort	<- you can use serach operators like AND OR as well as from: and to: eg '#JobsNow AND from:BarackObama' (without quotes)

4. Make a one off collection with TAGS > Run now! or set a trigger to collect every hour TAGS > Update archive every hour. To change the frequency open Tools -> Script Editor then Triggers -> Current script's triggers... and adjust

Advanced Settings:

Period	default	
Follower count filter	0	<- if search term is being spammed you can set the minimum followers a person must have to be included in archive
Number of tweets	10000	<- maximum varies based on the type of archive you are collecting
Type	search/tweets	<- use a search term in step 3 above to get results from last 7 days

Stats

Number of Tweets	3,707
Unique tweets	3,707
First Tweet	13/01/2013 20:15:00
Last Tweet	13/01/2013 21:45:00

Figure 10: Screenshot of the TAGS configuration spreadsheet.

After the creation of the coding scheme, we tagged all of the Tweets in the corpus with the 17 categories. This was done by simply adding a new column in the spreadsheet under the category's name.

Analysis / Visualization

Next, we analyzed the annotated data in order to gain insight into the vocabulary and the function of Tweets published about *Tatort*. Above all, we sought answers to the following questions:

(1) What functions of Tweets (with regard to the annotated category) are most frequent, i.e. do people primarily speculate about the potential murderer, or do they mostly joke about the plot?
(2) How does the function of Tweets relate to the time structure of the episode, i.e. will speculative Tweets decrease

throughout the course of the episode, as the plot unfolds?

(3) Are there typical words being used for different types of Tweets, i.e. will critique Tweets contain sentiment words such as "hate", "bad" or "terrible"?

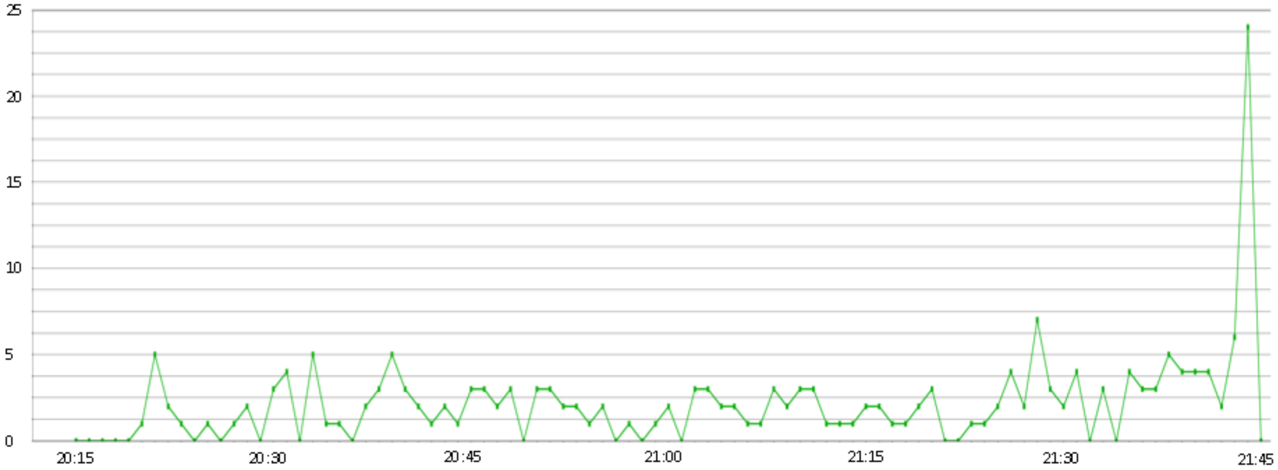
The results of the analysis are described in more detail in Burghardt et al. (2013), and thus will not be replicated in full length in this chapter. I will, however, provide the key insights, and will show how the analyses were undertaken and which tools were used for the visualization of the results.

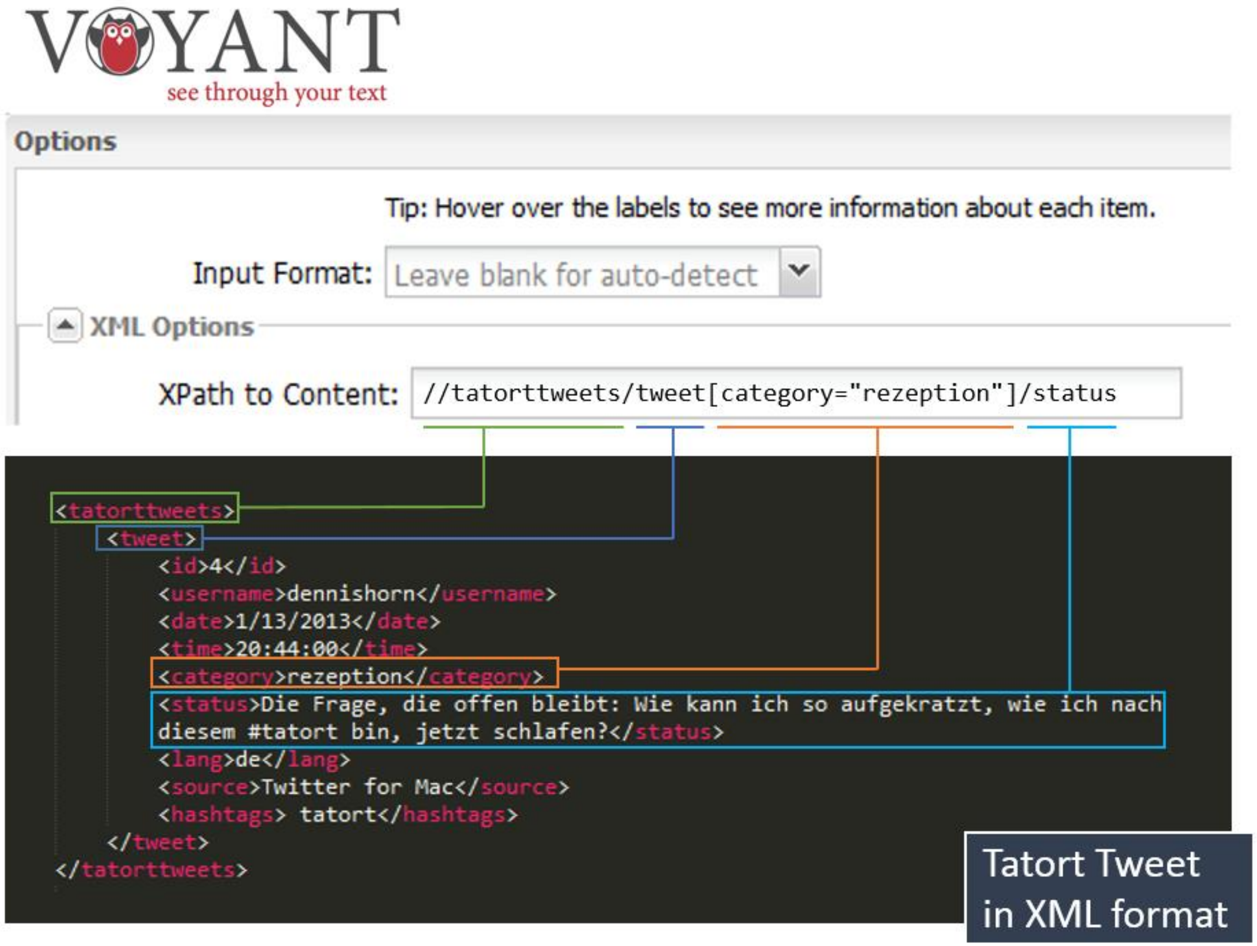
Basic functions of Tweets – The first question can be answered by simply counting the number of categories in Excel. By far the most Tweets were not original Tweets, but rather retweets of #tatort Tweets (19%) by other users. 10% of the Tweets described the reception situation (e.g. "lying in my bed and watching *Tatort* with my cat"), another 10% of the Tweets commented on the plot. Other types of Tweets were associations (7%), jokes (7%), comments about characters (6%), comments about dialogs (6%), critique (5%), speculation (5%), relation to another Tweet (5%), relation to *Tatort* in general (4%), logical flaws in the plot (3%), film production (3%), intermedial relations (3%), information about the start of the episode (2%), quotes from character dialogs (2%) and relations to social issues (1%).

Tweet functions and time – For the second question, the spreadsheet data was exported into the hierarchical XML document format, which contains information about the Tweet category and its exact time of publication. We used the eXist software and a number of corresponding XQuery (Boag et al. 2010) commands to analyze the corpus with regard to the frequency of Tweets of different categories. The software Altova XMLSpy was used to create diagrams that visualize the frequency along the time axis. Figure 11 shows the development of Tweets from the category critique, which shows a peak toward the end of the episode, indicating that people seem to wait with their critique until they have seen the whole episode.

Tweet functions and vocabulary – In order to answer the third question, we analyzed the XML corpus by using the online text analysis tool Voyant. Voyant allows users to specify sub-corpora within one larger corpus by means of an XPath (Clark & DeRose 1999) expression, i.e. it is possible to independently analyze the vocabulary for Tweets from the 17 different categories (Figure 12).

Figure 11: Frequency distribution of Tweets from the category critique throughout the course of the episode.





VOYANT
see through your text
Options
Tip: Hover over the labels to see more information about each item.
Input Format: Leave blank for auto-detect
XML Options
Tatort Tweet in XML format

XPath to Content: //tatorttweets/tweet[category="rezeption"]/status

```
<tatorttweets>
    <tweet>
        <id>4</id>
        <username>dennishorn</username>
        <date>1/13/2013</date>
        <time>20:44:00</time>
        <category>rezeption</category>
        <status>Die Frage, die offen bleibt: Wie kann ich so aufgekratzt, wie ich nach
        diesem #tatort bin, jetzt schlafen?</status>
        <lang>de</lang>
        <source>Twitter for Mac</source>
        <hashtags> tatort</hashtags>
    </tweet>
</tatorttweets>
```

Figure 12: XPath command that can be entered into Voyant to select the message text (status) of all Tweets that were tagged with category reception.

Voyant provides stop-word lists for different languages that can be used to filter highly frequent articles, prepositions, and conjunctions. The results of the frequency analyses are then visualized as an interactive word cloud. Figure 13 shows the most frequent words for Tweets from the category "comment" on the plot. Among the most frequent words are character names "Kopper" and "Brenner", but also "Hund" (reference to the death of a dog) and "Pink Floyd" (reference to commissar Kopper performing a Pink Floyd song).

6. Summary

This article has shown that a large amount of research is being dedicated to Twitter and the data that is produced by its users. From a corpus linguistic as well as from a media linguistic perspective, Tweets are promising objects of research, as they not only contain user generated language in the actual message text, but also a number of interesting metadata such as date, language, and location.

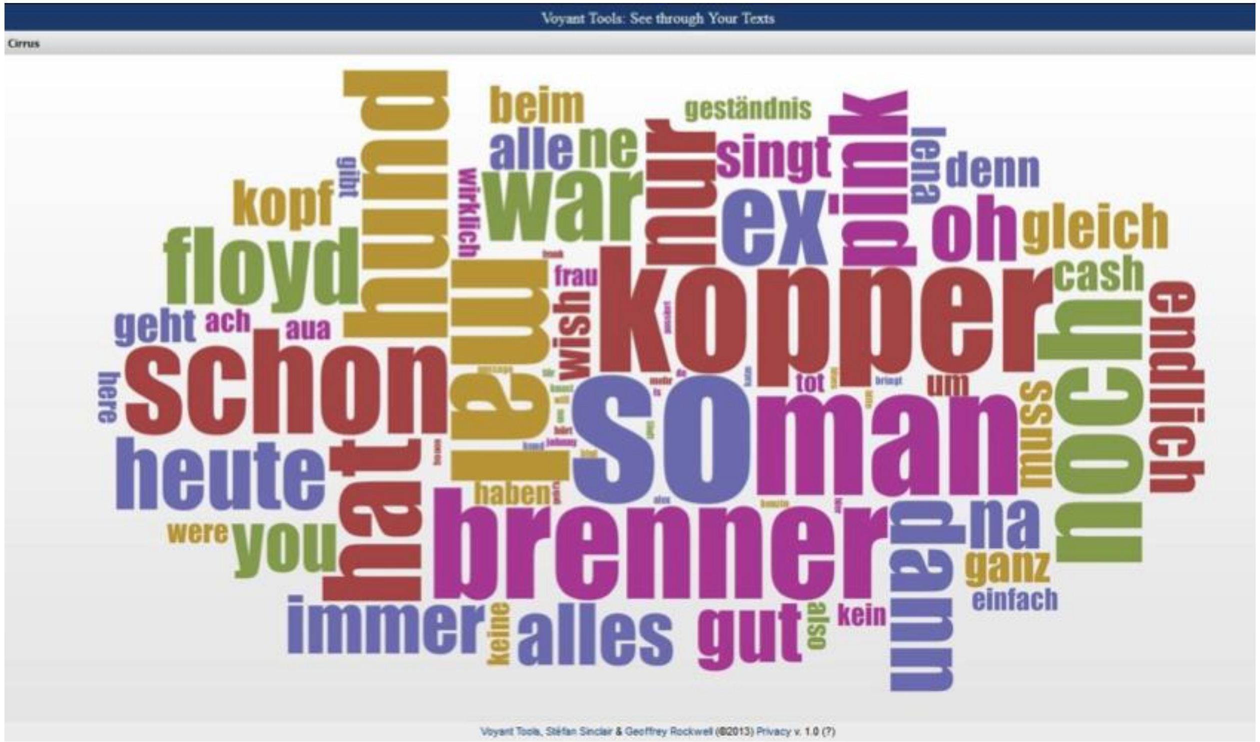


Figure 13: Word cloud for the most frequent words in Tweets that comment on the plot (visualization: Voyant Tools Version 1.0, Sinclair, S. & Rockwell, G., March, 2013).

Although Tweets are suited for a wide range of research studies, scholars often struggle to access Twitter data due to some limiting factors. Tweets are typically accessed via the official Twitter API, which requires basic programming skills that may not be available for scholars from the humanities (Burghardt & Wolff 2015). The API also comes with various rate limits, i.e. it is not possible to obtain arbitrary amounts of Tweets or Tweets that are older than one week. At the same time, Twitter's terms of use do not allow developers to pre-compile corpora of Tweets and share them with researchers outside of the Twitter platform.

In order to overcome the technical hurdles of the original Twitter API, a number of ready-to-use tools were introduced that can be used to collect, analyze and visualize Tweets. As most of the research on Twitter so far has been dedicated to aspects of communication structures, information behavior and various topics from the computer linguistics sector, this article aims at promoting more research from the media linguistics field by providing a basic introduction to available tools and methods.

References

Ausserhofer, J. & Maireder, A. (2012). National Politics on Twitter: Structures and Topics of a Networked Public Sphere. *Journal of Information, Communication & Society*, 16.3, pp. 291-314.

Baroni, M. et al. (2009). The Wacky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43.3, pp. 209-226.

Bazo, A. et al. (2013). TWORPUS – An Easy-to-Use Tool for the Creation of Tailored Twitter Corpora. In I. Gurevych et al. (Eds.), *Proceedings of the 25th International Conference of the German Society for Computational Linguistics and Language Technology*, *GSCL '13* (pp. 23-34). Heidelberg: Springer.

Beißwenger, M. & Storrer, A. (2008). Corpora of Computer-Mediated Communication. In A. Lüdeling & K. Merja (Eds.), *Corpus Linguistics: An International Handbook* (pp. 292-309). Berlin, New York: Mouton de Gruyter.

Boag, S. et al. (2010). XQuery 1.0: An XML Query Language (Second Edition) – W3C Recommendation 14 December 2010. Retrieved from http:/www.w3.org/TR/xquery/.

boyd, d. et al. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences* (pp. 1-10).

Burger, H. & Luginbühl, M. (2014). *Mediensprache: Eine Einführung in Sprache und Kommunikationsformen der Massenmedien*. Berlin & Boston: DeGruyter.

Burghardt, M. et al. (2013). Twitter als interaktive Erweiterung des Mediums Fernsehen: Inhaltliche Analyse von Tatort-Tweets. In *Workshop proceedings of the GSCL 2013*. Retrieved from http://tinyurl.com/Burgha2013.

Burghardt, M. & Wolff, C. (2015). Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik. In *Book of Abstracts Workshop "Informatik und die Digital Humanities", Leipzig*. Retrieved from http://tinyurl.com/burghaandwolff.

Clark, J. & DeRose, S. (1999, November 16). XML Path Language (XPath), Version 1.0, W3C Recommendation. World Wide Web Consortium (W3C). Retrieved from http://www.w3.org/TR/xpath/.

Crystal, D. (2006). *Language and the Internet*. Cambridge et al.: CUP.

Cunha, E. et al. (2011). Analyzing the Dynamic Evolution of Hashtags on Twitter: a Language-Based Approach. In *Proceedings of the Workshop on Language in Social Media LSM 2011* (pp. 58-65).

Derczynski, L. et al. (2013). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, ACL* (pp. 198-206).

Gimpel, K. et al. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 42-47).

González-Ibáñez, R. et al. (2011). Identifying Sarcasm in Twitter: a Closer Look. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers* (pp. 581-586).

Han, B. & Baldwin, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a # Twitter. In *Proc. of the 49th Annual Meeting of the ACL* (pp. 368-378). Retrieved from http://www.aclweb.org/anthology/P11-1038.

Jansen, B. J. et al. (2009). Twitter Power: Tweets as Electronic Word of Mouth. *Journal of the American Society for Information Science and Technology*, 60.11, pp. 2169-2188.

Java, A. et al. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (pp. 56-65).

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities)*. University of Illinois Press.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29.3, pp. 333-347.

Kricfalusi, E. (2015). The Twitter Hashtag: What Is It and How Do You Use It? Retrieved from http://techforluddites.com/the-twitter-hashtag-what-is-it-and-how-do-you-use-it/.

Kumar, S. et al. (2013). *Twitter Data Analytics*. New York: Springer.

Larkin, J. H. & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, 11.1, pp. 65-100.

Makice, K. (2009). *Twitter API: Up and Running*. Beijing et al.: O'Reilly.

Manovich, L. (2001). *The Language of New Media*. Cambridge (MA) & London (UK): MIT Press.

Marx, K. & Weidacher, G. (2014). *Internetlinguistik: Ein Lehr- und Arbeitsbuch*. Tübingen: Narr Verlag.

Mazza, R. (2009). *Introduction to Information Visualization*. London: Springer.

McCreadie, R. et al. (2012). On Building a Reusable Twitter Corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 12* (pp. 1113-1114).

Meier, F. & Elsweiler, D. (2014). Personal Information Management and Social Networks Re-finding on Twitter. In *Proceedings of the 5th Information Interaction in Context Symposium* (pp. 339-341).

Meier, F. et al. (2014). More than Liking and Bookmarking? Towards Understanding Twitter Favouriting Behaviour. In *Proc. of the 8th International Conference on Weblogs and Social Media (ICWSM)*. Retrieved from http://www.cs.nott.ac.uk/~mlw/pubs/icwsm2014-favouriting.pdf.

Michel, J.-B. et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331.6014, pp. 176-182.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Overbeck, A. (2014). Twitterdämmerung: ein textlinguistischer Klassifikationsversuch. In N. Rentel et al. (Eds.), *Von der Zeitung zur Twitterdämmerung* (pp. 207-228). Berlin: LIT Verlag.

Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the LREC* (pp. 1320-1326).

Proulx, M. & Shepatin, S. (2012). *Social TV: How Marketers Can Reach and Engage Audiences by Connecting Television to the Web, Social Media, and Mobile*. Wiley & Sons.

Rehbein, I. (2013). Fine-grained POS Tagging of German Tweets. In *Proceedings of the 25th International Conference of the German Society for Computational Linguistics and Language Technology, GSCL '13* (pp. 162-175).

Rogers, R. (2013). Debanalizing Twitter: The Transformation of an Object of Study. In *Proc. 5th Annual ACM Web Science Conference* (pp. 356-365).

Ross, C. et al. (2011). Enabled Backchannel: Conference Twitter Use by Digital Humanists. *Journal of Documentation*, 67.2, 214-237.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. 2nd edition. Sebastopol CA: O'Reilly.

Spanner, S. et al. (2015). Twista – An Application for the Analysis and Visualization of Tailored Tweet Collections. In *Proceedings of the 14th International Symposium of Information Science (ISI 2015)* (pp. 191-202).

Wagner, K. (2013). You Can Now Search for Any Tweet in History. Mashable.com. Retrieved from http://mashable.com/2013/09/04/search-tweet-history/.

Williams, S. A. et al. (2013). What Do People Study When They Study Twitter? Classifying Twitter Related Academic Papers. *Journal of Documentation*, 69.3, 384-410.

Zanzotto, F. M. et al. (2011). Linguistic Redundancy in Twitter. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing, ACL* (pp. 659-669).