Using Social Media Data for Research: An Overview of Tools

Dr. Wasim Ahmed

Doctoral Graduate

Information School

University of Sheffield

Correspondence:

+44 749 1655 862

wahmed1@ sheffield.ac.uk

Original manuscript accepted for publication in

Journal of Communication Technology

Published by the Communication Technology Division of the Association for Education in Journalism and Mass Communication

Using Social Media Data for Research: An Overview of Tools

Social media platforms have grown in popularity with the use of social media websites increasing over recent years as more households, organisations, and individuals have access to the Internet (OECD, 2016). Today there are more social media platforms than ever with more members of the public using social media, and more citizens, businesses, as well as charities and other organisations using social media platforms (Chaffey, 2016).

Social media data are increasingly becoming primary sources of data for social research. Our day to day interactions and our personal and professional relationships can be said to be made up via online interactions such as posts, comments, favourites, tags, likes, and links. All of our interactions leave traces in the form of data which can be analysed for research purposes.

Those who lack a programming, and/or a technical background may at first feel daunted when first considering how to retrieve social media data. However, now more than ever, there are a plethora of research tools that can be used by those from the social sciences. This article builds, expands, and aims to consolidate previous published practical guides (Bruns & Liang, 2012; Ahmed, 2015a; Ahmed 2015b; Ahmed & Thomson, 2016) as well as tutorials and seminars that look at specific applications (Ahmed, 2016a; Ahmed, 2016b; Ahmed, 2016c; Ahmed, 2016d).

The aim of this article is to provide a succinct overview of the range of social media tools that can retrieve data from platforms such as: Twitter, Facebook, Instagram, YouTube, Flickr, Foursquare, Panoramio, AIS Shipping, news blogs, and VK among others. Each of the applications outlined have been used and tested by the author, to include the commercial applications. As with all academic research involving human participants, research using social media data should consider all possible ethical implications and research should be conducted with rigour.

The article hopes to act as a valuable resource for social media researchers and will highlight the different types of quantitative and qualitative social media analysis techniques associated with each tool. These range from the analysis of both text and image data, network analysis, geographical analysis, and more. No previous empirical research has tested and/or presented this amount of web-based and/or desktop applications that can be used to retrieve data from Twitter. Tools are reported below, in alphabetical order, and Table 1 provides an overview of the different applications.

2. Overview of Tools

2.1 Boston University Twitter Collection and Analysis Toolkit (BU-TCAT)

The Boston University Twitter Collection and Analysis Toolkit (BUT-TCAT, n.d.) is a web-based tool that allows users to retrieve Twitter data via the Stream Application Programming Interface (API) (i.e., the Gardenhose). This tool provides ability to perform a number of metrics, and data can be sorted by algorithms, which enables researchers to find people and/or importance. Once data is retrieved it can be processed for analysis, and users are able to perform some of the following:

- Examine Twitter timelines, with minute by minute timestamping
- The ability to generate tweet statistics such as hashtag, retweet, geocoding, and unique users
- Activity metrics such as user visibility and mention frequency
- Hashtag frequency, such as hashtag-user activity and word and identical tweet frequency
- Lists of individual retweets and geocoded tweets
- Network graphs by mentions, co-hashtagging, and hashtag-user graphs
- Cascades, alluvial diagrams, and associational profiles

- Once a query is entered into the BU-TCAT, it retrieves data continuously until they are requested to be turned off.

The BU-TCAT can be accessed only via request by academic researchers, free of charge, and only for non-commercial purposes. A valid institutional email address is required in order to request access.

2.2 Chorus Project

Chorus (Brooker, Barnett, & Cribbin, 2016), retrieves Twitter data from Twitter's Search API. A software development project that began around 2011, Chorus traces its origins to Brunel University and specifically the *MATCH* and *FoodRisC projects.* The Chorus Project is comprised of two programs: i) Tweetcatcher, which retrieves Twitter data, and ii) Tweetvis, which analyses Twitter data. Once data is retrieved in Chorus it is possible to do the following:

- Visualize occurrence of tweets over time in the Timeline Explorer
- Analyse Twitter conversations according to a number of metrics such as: tweet frequency, sentiment, semantic novelty, and homogeneity and collocated words, among others.
- Visualize tweets using the Cluster Explorer

Chorus can be downloaded by completing a download form, and software access can be requested by those from any sector.

2.3 COSMOS Project

The Collaborative Online Social Media Observatory (COSMOS) application was a collaborative project that ran from 2012 to 2015, and the project is now continued and the software maintained by the ESRC-funded Social Data Science Lab (Burnap et al., 2014). COSMO can retrieve Twitter data from either the Search or Stream APIs. Once data is retrieved it can be processed for analysis, and users are able to perform some of the following:

- Generate word clouds
- Frequency charts
- Network graphs
- Ability to plot longitude and latitude data on a map

COSMO works best on Mac OS X and Linux, and the developers of the application recommend using Windows only as a last resort. COSMO can be accessed only via request by academic researchers, free of charge, and only for non-commercial purposes. A valid institutional email address is required in order to request access.

2.4 DiscoverText

DiscoverText (n.d.) is a licensed reseller of Twitter data. Via Texifter, this online application can import tweets from the global stream of Twitter data known as the *Firehose API*. Once data is retrieved, it can be processed for analysis, and users are able to perform some of the following:

- Filter data using a number of metrics
- De-duplicate data and find near duplicate clusters
- Cluster and search data
- Human code data, and/or
- Machine classify large numbers of small unstructured units of text
- Take random samples of data for coding
- Visualize tweets using time series graph
- Produce metrics using the Meta feature
- Examine number of links, retweets (RTs), number of tweets, Reply IDs, In Reply @s, and tweet rate

DiscoverText offers a free, 3-day trial, and the platform can handle social media data from a number of platforms such as Twitter, Facebook, blogs, forums, and online news platforms. DiscoverText also allows users to import data into the platform, and there are also features which allow users to work in collaboration.

2.5 Echosec

Echosec (n.d.) allows end users to navigate to any location in the world and examine the social media activity around the area. Currently Echosec Pro allows users to access at least the following social data feeds: Instagram, Twitter, Foursquare, Panoramio, AIS Shipping, Sina Weibo, Flickr, YouTube, and VK. Users are able to perform some of the following:

- Search locations for social media content
- Search for topics of interest and examine words associated with geographical areas
- Ability to search by username
- Set up custom alerts

Echosec works by making use of location-based metadata to search for social media and other open source information. It relies mostly on a range of API requests directly to the social media networks (Twitter, Instagram, Facebook, etc.), but also does rely on third party information repositories. There is a free version of Echosec, and prices range from $89 to $129 per month.

2.6 Follow the Hashtag

Follow the Hashtag (n.d.) is an online commercial application that has access to Twitter's Search and Firehose APIs. However, the application also allows users to retrieve and analyse a limited number of tweets for free. Once data has been retrieved it is possible to:

- Discover influential users
- Perform gender analysis

- Produce visualisations such as time series graphs and blob charts
- Plot locations of users on a map
- Perform sentiment analysis as well as deep search

The free version limits users to a maximum of n=1500 tweets going back in time seven days, and a n=500 limit on the amount of tweets that can be exported.

2.7 Mozdeh

A free Windows application, Mozdeh (n.d.) is a creation of the Statistical Cybermetrics Research Group at the University of Wolverhampton that was created in conjunction with the *Creen,* and *CyberEmotions* EU projects. Once data is retrieved from Twitter it is possible to:

- Search tweets
- Find gender differences in tweets
- Draw time series graphs
- Conduct word frequency analyses
- Create networks of users from their tweets
- Save a random sample of tweets for content analysis
- Remove spam tweets

Mozdeh allows users to download data via Twitter's Search API, and it is recommended that users install and use the application on a desktop computer.

2.8 Netlytic

Netlytic (n.d.) is cross platform, web-based tool that can be used to retrieve and analyse data from Twitter, Facebook, YouTube, Instagram, and RSS Feeds. Netlytic has two tiers which are available for free and can handle a maximum of n=10000 records. Netlytic allows users to do the following:

- Capture data from social media websites

- Discover popular topics
- Find and explore emerging themes of discussion
- Build, visualize, and analyse data using social network analysis
- Ability to plot longitude and latitude data on a map

Netlytic is part of the *The Social Media Lab.*

2.9 NodeXL

NodeXL (n.d.) is a Microsoft Excel plugin. The software can be used to obtain data from Twitter, YouTube, and Flicker. NodeXL runs on Windows operating systems and is supported by the Social Media Research Foundation. Once data is retrieved NodeXL, users can process the data in order to do the following:

- Visualize extracted data based on a number of graph layout algorithms
- Create a time series graph
- Examine the most frequently shared URLs, domains, hashtags, words, word pairs, replied-tos, mentioned users
- Produce metrics overall and by group level

NodeXL has been used for academic and commercial research. There is a free version of NodeXL, as well as a *Student* and *Pro Version*.

2.10 Twitter Arching Google Spreadsheet

TAGS (Twitter Arching Google Spreadsheet) (n.d.) is a web-based tool that consequently works across a number of operating systems (TAGS, n.d.). Using TAGS, which accesses the Twitter's Search API, researchers may retrieve data from Twitter. After capturing Twitter data, it is possible to use TAGS to do the following:

- View metrics such as number of links, RTs, number of tweets, unique tweets, in Reply IDs, in Reply @s, and tweet rate

- Explore tweets within the TAGS Explorer
- Search tweets

Similar to the BU-TCAT and DiscoverText, TAGS allows users to set up automatic retrieval at varying, specified time intervals. As it is cloud based, TAGS does not require a computer to run continuously in order to retrieve tweets.

2.11 Twitonomy

Twitonomy (n.d.), which uses the Search API, is a web-based Twitter analytics tool that has both a free and premium version. Once a user has searched for a keyword and/or a user and retrieved data, that user may analyse the data by doing the following:

- Generating visual analytics such on tweets, retweets, replies, mentions, and hashtags
- Produce insights into Twitter users such as: tweets per day, retweets, user mentions, replies, links, hashtags, tweets retweeted, and tweets favourited.

The free version of Twitonomy allows unlimited analysis of user accounts. However, the ability to search for keywords requires users to upgrade to a premium version, which starts at $19 a month or $199 a year.

2.12 NVivo

NVivo (QSR International, n.d.) is used within the social sciences for analysing qualitative data. However, NVivo also provides some really useful “non-programming” approaches to analysing social media data. Using the NVivo NCapture plugin, users can retrieve Twitter data and apply a number of analysis techniques such as the following:

- Filter attributes based on username or number of followers
- Auto-code dataset based on a number of attributes
- Analyse sentiment in the Twitter dataset and create a sociogram (NVivo 11 Plus for Windows)

In NVivo users can import data from other sources. For example, if users had already retrieved Twitter, Facebook or other web-based data using NVivo, then it would be possible to import this data into the application.

2.13 Pulsar Social

Pulsar Social(n.d.) allows users to retrieve and analyse social media data from a number of platforms such as Twitter, Facebook, Instagram, and blogs. Pulsar is web based, and prices range from £500/$700 per month, which allows unlimited searches and the ability to retrieve 100 thousand “interactions.” Once a user has retrieved data, Pulsar enables users to do the following:

- Perform image and video mining
- Manipulate, explore, and interact with data
- Visualize data
- Run sentiment analysis
- Plot tweeter locations on a map
- Filter data

Pulsar users can also set up a query to retrieve and analyse data from a number of different social media platforms.

2.14 SocioViz

SocioViz (n.d.) is a free online social media analytics application that is powered by social network analysis metrics using Twitter’s Search API, and can be used to retrieve data from Twitter. Once data is retrieved users can do the following:

- Analyse specific topics, terms, and hashtags,
- Identify key influencers, opinions, and contents
- Export data

SocioViz is a non-profit organisation, and users are required to complete a short form in order to register.

2.15 Visibrain

Visibrain (n.d.) is a web-based Twitter monitoring platform for digital marketing professions, and has access to Twitter's Firehose API. Once data has been retrieved by Visibrain, users can analyse the data by the following techniques:

- Creating time series graphs
- Examining metrics such as most frequently occurring keywords, hashtags, users, locations, tweets, retweets
- Examine the type of audience tweeting, device tweeting from, and content types
- Export data into Gephi for visualization

There are different tiers of Visibrain such as the *mini tier* which costs £300/$428 a month and has the capability of retrieving n=10,000 *mentions*.

2.16 Webometric Analyst

Webometric Analyst (n.d.) can retrieve general and specific types of data from a number of online sources such as YouTube, Twitter, Tumblr, Flicker, and Mendeley, as well as other web sources. Once a user has retrieved data, Webometric can do the following:

- Create network diagrams
- Create Impact reports
- Extract images from Tweets and/or Flicker posts

In order to use Webometric Analyst, users are required to obtain a key (a character string) from the Windows Azure market place, and a free key allows up to 5,000 searchers per month. Webometric Analyst has the ability to repair data from a previously outlined application

Mozdeh. That is, if Mozdeh crashes either due to a power cut or due to a fault in the computer, Webometric Analyst has the ability to completely repair the files.

2.17 Other

Other applications are available, but this author has not tested them. However, they have been provided here in an effort to provide a complete list of resources available to social media analysts. These applications include the following:

- DMI-TCAT[1] (free)
- Crimson Hexagon[2] (commercial)
- Brandwatch[3] (commercial)

Those interested in Twitter data collection tools including modules and libraries that require programming knowledge may also be interested in Deen Freelon's curated list of tools.[4] Additionally, a number of advance data analysis and statistical applications can be used to analyse social media data, including the following:

- R[5]
- SPSS[6]
- KNIME[7]
- Weka[8]
- Tableau[9]

[1] https://github.com/digitalmethodsinitiative/dmi-tcat
[2] http://www.crimsonhexagon.com/
[3] http://www.brandwatch.com/
[4] http://socialmediadata.wikidot.com/
[5] https://www.r-project.org/
[6] http://www.ibm.com/analytics/us/en/technology/spss/
[7] https://www.knime.org/
[8] http://www.cs.waikato.ac.nz/ml/weka/
[9] http://www.tableau.com/

These applications should also be researched when deciding which application should be selected for a project.

Table 1 – Overview of tested tools that can be used to retrieve and analyse social media data

Tool	**OS**	**Download from**	**Platforms***
Chorus	Windows (Desktop advisable)	http://chorusanalytics.co.uk/chorus/request_download.php	Twitter
COSMOS Project	Windows MAC OS X	http://www.cosmosproject.net/	Twitter
DiscoverText (free 3 day trial)**	Web-based	http://discovertext.com/	Twitter Facebook Blogs Forums Online news platforms
Echosec	Web-based	https://www.echosec.net/	Instagram Twitter Foursquare Panoramio AIS Shipping Sina Weibo Flickr YouTube VK
Follow the hashtag	Web-based	http://www.followthehashtag.com/	Twitter
Mozdeh	Windows (Desktop advisable)	http://mozdeh.wlv.ac.uk/installation.html	Twitter

Netlytic	Web-based	https://netlytic.org/	Twitter Facebook YouTube Instagram RSS Feed
NodeXL	Windows	http://nodexl.codeplex.com/	Twitter YouTube Flicker Facebook
Twitter Arching Google Spreadsheet (TAGS)	Web-based	https://tags.hawksey.info/	Twitter
Twitonomy	Web-based	http://www.twitonomy.com/	Twitter
Visibrain	Web-based	http://www.visibrain.com/en/	Twitter
Nvivo	Windows MAC		Twitter*
PULSAR Social	Web-based	http://www.pulsarplatform.com/	
Socioviz	Web-based	http://socioviz.net/	Twitter
Webometric Analyst	Windows	http://lexiurl.wlv.ac.uk/	Twitter (with image extraction capabilities) YouTube Flickr Mendeley Other web resources

*Some applications such as NVivo and DiscoverText allow users to import data from a variety of data sources

3. Summary

Each of the desktop and web-based software application outlined above have been tested by the author, and at the time of writing were all active and functional. The majority of the applications identified predominately handle Twitter data. Looking forward in the area of software development, it will be valuable for the field of social media if developers look to build applications that could obtain data from other social media platforms such as Pinterest, Goolge+, Tumblr, Instagram, Flickr, Vine, LinkedIn, and Amazon among others. Moreover, at the Masters and Ph.D. level, more emphasis should be placed on training for social science students in effectively using existing software that captures data analyse data from social media platforms.

References

Ahmed, W. (2016a, November). Insights from Social Media. In A, Fenton (Chair), Symposium conducted at the meeting Creative Entrepreneur, Media City, Salford.

Ahmed, W. (2016b, August) Social Media Analytics. In A, Grace (Chair), Symposium conducted at the meeting Social Media Research Seminars Department for Work and Pensions (DWP). London.

Ahmed, W. (2016c, June) workshop on Twitter Analytics. In Lugović, S (Chair), Symposium conducted at the meeting the Contemporary Issues In Economy & Technology (CIET). Spit.

Ahmed, W. (2016d, October). Introduction to NodeXL. In Jessop, C (Chair), Symposium conducted at the meetingAn Introduction to Tools for Social Media Research. London

Ahmed, W. (2015a) Using Twitter as a data source: An overview of current social media research tools LSE Impact of Social Sciences blog. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/10/social-media-research-tools-overview/ (Accessed 11/07/2015).

Ahmed, W., Thomson, S. (2016, September). Twitter and crisis communication: an overview of tools for handling social media in real time [Web log post]. *LSE Impact of Social Sciences Blog*. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2015/09/28/challenges-of-using-twitter-as-a-data-source-resources/

Ahmed, W. (2015b, January). A list of tools to capture data from Twitter [Web log post]. *Wasim Ahmed: A blog about software, analytics, tricks, tips and occasional opinion*. Retrieved from https://wasimahmed.org/2015/01/30/a-list-of-tools-to-capture-twitter-data/

Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society, 3*(2).

Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J., . . . & Conejero, J. (2014). COSMOS: Towards an Integrated and Scalable Service for Analyzing Social Media on Demand. *International Journal of Parallel, Emergent and Distributed Systems. 30(2), 80-100.*

BU-TCAT. (n.d.). Twitter Collection and Analysis Toolkit (TCAT) at Boston University. Retrieved from http://www.bu.edu/com/research/bu-tcat/

DiscoverText. (n.d.). Retrieved from https://www.discovertext.com/

Echosec. (n.d.). Retrieved from https://www.echosec.net/

Followthehashtag. (n.d.). Retrieved from: http://www.followthehashtag.com/

TAGS (n.d.). Retrieved from: https://tags.hawksey.info/.

Statistical Cybermetrics Research Group. (n.d.). *Mozdeh Twitter Time Series Analysis*. Retrieved from http://mozdeh.wlv.ac.uk/

Social Media Research Foundation. (n.d.). *NodeXL*. Retrieved from http://www.smrfoundation.org/nodexl/

Chaffey, D. (2016). Global social media research summary. *Smart Insights: Social Media Marketing*. *13-32*.

SocioViz. (n.d.). Retrieved from http://socioviz.net/SNA/eu/sna/login.jsp

Twitonomy. (n.d.). Retrieved from https://www.twitonomy.com/

OECD. (2016). [Online]. Internet access (indicator). Retrieved from https://data.oecd.org/ict/internet-access.htm. doi: 10.1787/69c2b997-en

Pulsar. (n.d.). Retrieved from http://www.pulsarplatform.com/

QSR International. (n.d.). Retrieved from http://www.qsrinternational.com/product

Visibrain. (n.d.). Retrieved from http://www.visibrain.com/en/

Webometric Analyst. (n.d.). Retrieved from http://lexiurl.wlv.ac.uk