

Sentiment Analysis and Time Series with Twitter

Mike Thelwall, School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

E-mail: m.thelwall@wlv.ac.uk. Tel: +44 1902 321470 Fax: +44 1902 321478.

Introduction

The contributions of ordinary members of the public to microblogging services like Twitter and Weibo can give social scientists unique insights into public reactions to specific events and to changes in public opinion over time. For example, the changing volume of tweeting around an event, such as the UK riots of 2011 or a United Nations Conference on Sustainable Development, can reveal when the public first became interested in it and when this interest started to fade. Peaks in the volume of tweeting can also point to which instances within a broad event generated the most interest. Both blogs and microblogs are particularly useful for monitoring public opinion in this way because (a) they are reliably time-stamped, unlike most of the rest of the web, so that they can be analysed from a temporal perspective, (b) they are relatively easy to create, so that a wide segment of the population with internet access could, in theory, create them, and (c) they are public and hence accessible to researchers, unlike most social network sites. Microblogs are probably created by a wider segment of the population than blogs in most countries and are naturally updated more frequently, making them a better source of public opinion information than blogs for many purposes.

A number of previous studies have analysed Twitter or blogs over time. An early study of blogging relating to the Danish Cartoons affair showed that there was almost no blogging about it when it first happened and that it became a major news story because of two events that occurred five months after the cartoons were published, rather than the publication of the cartoons themselves (Thelwall, 2007). The advantage of blogs for this study was to show that there was little interest in the topic, at least in the blogosphere, when the cartoons were published. Another study compared the events attracting the most Twitter attention in six different English-speaking countries, finding significant overlaps in interest that could be explained by geopolitical factors (D. Wilkinson & Thelwall, 2012). On a smaller scale, an investigation of 7,184 tweets relating to three different campus shootings was used to show how the medium was used to help make sense of the situation at different points in time (Heverin & Zach, 2012) Trends in tweeting have even been used to identify news about real-world events (Becker, Naaman, & Gravano, 2011).

The time series approach can also be used to investigate changes in sentiment over time, either to understand the role of sentiment in an event or changes in popularity over time. One large-scale study compared overall changes in sentiment in tweets over time with external social, political, cultural and economic phenomena, finding a connection between offline events and online sentiment (Bollen, Pepe, & Mao, 2011). A number of studies have focused on sentiment in relation to elections to assess whether it is possible to predict outcomes (Chung & Mustafaraj, 2011). Whilst it seems logical that sentiment expressed in Twitter would reflect the public mood, there is a problem with Spam, attempts to manipulate Twitter for political goals, and different levels of internet use for people of differing political persuasions that makes this difficult in practice.

The remainder of this chapter describes how to conduct a time series and sentiment analysis of Twitter. It gives an overview of methodological issues and describes in broad detail how to use a specific set of software tools developed at the University of Wolverhampton in order to gather and analyse Tweets for a specific topic. More specific instructions about how to use the software described and additional examples of analyses can be found at <http://mozdeh.wlv.ac.uk/>.

Creating a corpus of tweets

At the time of writing the official Twitter search engine, search.twitter.com, seemed to only return tweets that were up to two weeks old. In addition there are currently no free archives of tweets and hence no convenient source of tweets for a Twitter time series analysis. Researchers interested in analysing large sets of tweets therefore need to either buy a relevant collection of tweets or monitor Twitter over a period of time in order to build their own corpus of tweets. This chapter describes the latter approach.

A simple way to create a corpus of tweets would be to search Twitter periodically, such as daily, recording and saving the results. For instance, the search might be a hashtag or a more complex query or set of queries designed to match topic-relevant tweets. This would be time consuming and would also be ineffective if there were too many matching queries to save. A practical alternative is to use a computer program to automatically submit the queries periodically and save the results. This approach is described below, following a brief discussion of ethical considerations and limitations.

Creating a corpus of tweets does not have the same ethical and privacy implications as interview transcripts or questionnaire data because tweets are inherently public and readable, when posted, by anyone with an internet connection. Hence, they should be regarded as documents rather than human-related data (D. Wilkinson & Thelwall, 2011). Researchers should nevertheless avoid republishing individual tweets or their corpus of Tweets as this could have privacy implications caused by drawing attention to the individuals concerned. For example, if investigating suicide-related tweets it could have negative consequences if a suicidal tweet was quoted in an article and the tweeter or their acquaintances found it.

In practical terms, the software Webometric Analyst can be used to gather Tweets (see <http://mozhdew.wlv.ac.uk/> for instructions). This program can automatically submit a pre-defined set of queries to Twitter every hour and record the results for later analysis. Webometric Analyst works by submitting the queries to Twitter via the officially permitted route, the Twitter API (Applications Programming Interface). At the time of writing, the Twitter API permitted keyword queries and allowed the query to specify the language of the tweets and the approximate geographic location of the tweeters. These features are useful to ensure that only the most relevant tweets are gathered. To use this facility, the following steps are recommended.

1. Construct and test a set of queries for the topic researched. As far as possible these queries should collectively give good coverage of the topic so that most tweets relevant to the topic would match at least one of the queries. As far as possible, the queries should not match tweets that are Spam or otherwise irrelevant to the topic. Normally the second consideration is most important for time series analyses because it is difficult to filter out the irrelevant tweets and they can result in irrelevant analysis results. The queries should be tested in search.twitter.com before being used and the first results from Webometric Analyst should also be checked for accuracy and appropriateness.
2. Run the tested queries in Webometric Analyst on a computer that is permanently switched on. This will collect tweets in real-time so the queries should be set up before, or shortly after the start of, the event to be monitored. Webometric Analyst should be left going for the duration of the event to be monitored or as long as makes sense for the analysis.
3. Process the tweets gathered using the methods below.

Gathering Tweets using the method above has a sampling limitation. Twitter may not return all tweets that match a query and may impose arbitrary restrictions, such as delivering only 10% of the matches. Hence the results should be treated as a sample rather than a

comprehensive collection. Twitter also returns a maximum number of results per query (currently 1,000) so a query may give especially incomplete results during busy periods.

A simple time series analysis of a Twitter corpus

A simple way to analyse temporal trends in a Twitter corpus would be to sample a specified number of tweets at different time periods, such as at the beginning, middle and end, and then use a content analysis (Neuendorf, 2002) to classify the samples. A comparison of the results at different time periods could then be used to identify changes over time. In contrast, the methods described in the remainder of this section are more quantitative and use a graphical approach to identify trends in volume of tweeting over time. Nevertheless, it is a good idea to use content analysis in conjunction with the graphical approach in order to get deeper and more qualitative insights into the data.

The graphical time series approach is essentially to construct a graph of the volume of topic-relevant tweeting over time and to use the shape of the graph to identify trends in interest in the topic as well as individual events of interest during the time monitored. The graph used is the number of tweets in the corpus (i.e., matching the set of queries used to generate the corpus) plotted against time. The time interval used is normally hours, so that each point on the graph represents the number of topic-relevant tweets gathered within a single hour. Figure 1 is an example of such a graph created from a corpus of tweets about the UK riots of 2011.

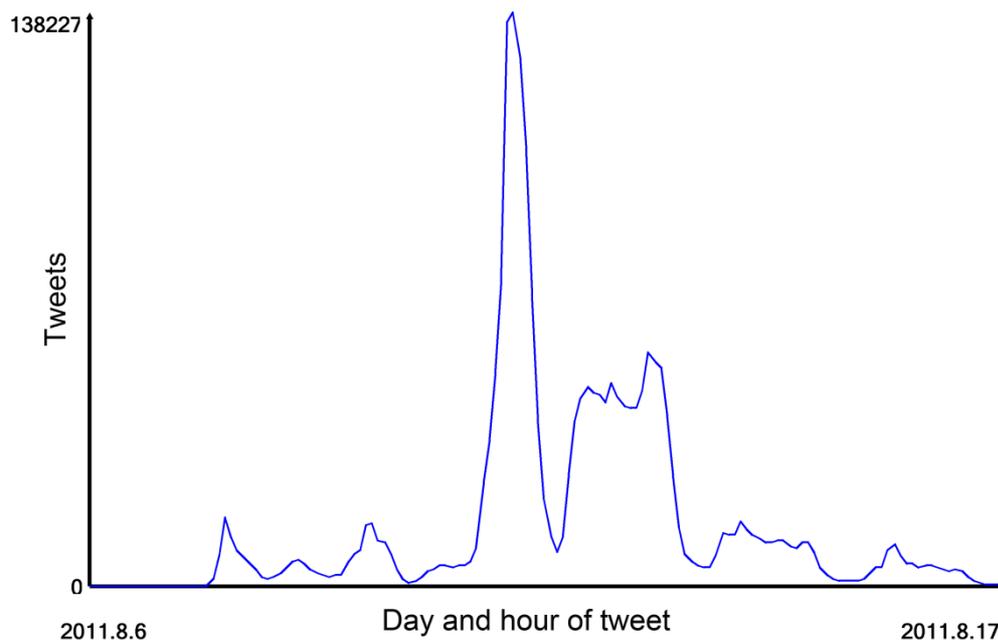


Figure 1. Volume of tweeting over time relating to the UK Riots of 2011. At the peak, at least 138,227 relevant tweets were sent in a single hour. It is clear from the graph that the peak in interest was considerably after the start of the riots. Note that some of the fluctuations are caused by time of day, with troughs during each night.

Visual inspection of the graph can be used to identify the overall trend in interest in the topic as well as key points of interest. If the graph is for a specific time-limited topic (e.g., a conference) then the following should be observed.

- The initial increase in volume of tweeting points to the time at which the topic started to gain interest
- The graph is likely to show a decrease in volume, pointing to the time at which people started to lose interest in the topic or ceased to find new information about it to post.

If the topic is a long running or permanent issue, such as interest in a political party then the following should be observed.

- Is the volume of interest increasing or decreasing over time?
- Is any increase or decrease approximately constant or are there changes in the broad pattern?

For either type of graph the following should be searched for.

- Spikes indicating specific events of interest for the topic. To detect what the event of interest is, a sample of tweets from the spike should be read. This can be conducted as a formal content analysis, if necessary. Spikes can sometimes be caused by Spam and so it is important to check that each spike analysed is genuine by checking the sample for suspicious content.

The identification of spikes can be difficult because twitter time series graphs are likely to be quite spikey due to natural variation in the data rather than due to external events. Hence, when detecting spikes, only the largest spikes should be investigated. Moreover, if a graph is very jagged so that it appears to be spikey without any trend then it should be redrawn with a longer time period (e.g., with each point corresponding to the number of tweets in a whole day rather than a whole hour) in order to get a less jagged line.

Time series analysis of queries within a Twitter corpus

The above analyses can be repeated for a subset of a corpus by constructing a graph for the tweets matching a given query. This can reveal patterns of interest for the subtopic represented by the query within the overall topic. The methods described above can be repeated except based on a graph of the *percentage* of tweets matching the query out of all tweets in the corpus. This is useful for tracking an issue or other theme throughout the corpus. Figure 2 shows how this works for the issue of the police. As is clear from the graph, the police were central to tweeting about the riots with 25% of tweets explicitly mentioning them. They were subsequently less frequently tweeted about but remained a significant topic of interest throughout. Graphs can also be constructed with more complex queries to capture issues that may be expressed with different words. For instance, to track gender issues in any topic then the query *male female man woman men women* could be submitted to capture and graph several different ways in which gender may be mentioned. It is primarily up to the researcher to produce a list of topics to analyse in this way and to convert them into effective queries to produce a graph.

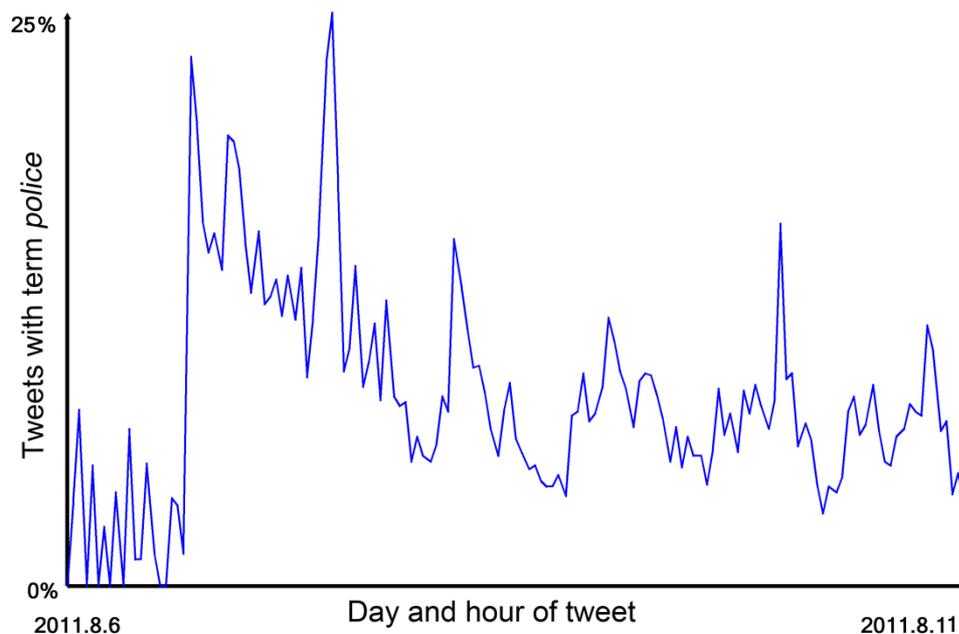


Figure 2. Percentage of tweets from a UK riots 2011 corpus that mentioned the term police. The graph shows a clear decrease in interest over time, although even at the lowest periods, 5-10% of tweets mentioned the police and there are periodic spikes in interest in the police.

A Twitter corpus can be converted into a time series graph using software. This can be much more efficient than using a manual approach – perhaps aided by a spreadsheet. The time series graphing and analysis program Mozdeh is described here. A Twitter corpus gathered can be converted into a graph by Mozdeh in a number of steps, as described fully online (<http://mozdeh.wlv.ac.uk>). First, the corpus must be converted into the file format used by Mozdeh and then Mozdeh has to index the tweets in order to create a graph. Once the indexing is complete, Mozdeh will create a time series graph of the data. This can then be used to identify the spikes or trends discussed above. Mozdeh can also produce the graphs for subsets of the tweets by entering a query with the graph so that it produces a graph of how frequently the tweets within the corpus match the query. The same steps described above can be used to analyse this graph. It makes sense to analyse keywords relevant to the broad topic to identify how they relate to the overall topic.

Mozdeh also has a more complex function that automatically identifies the 1000 individual words within the corpus that exhibit the biggest increase in frequency over time. This option can be used to automatically identify important events within the topic as an alternative to the spike method. This is because words associated with an increase in frequency within the corpus are normally associated with spikes of increases interest caused by a particular topic.

Sentiment analysis for Twitter

Automatic sentiment analysis has become popular over the past decade, especially for web data. A sentiment analysis program predicts the sentiment content of texts based upon features it identifies, such as the words used and the presence of emoticons. This section describes a sentiment analysis program designed for social web data, SentiStrength, and explains how it can contribute to analyses of a Twitter corpus. The availability of software to conduct sentiment analysis makes it possible to run large scale investigations into sentiment in Twitter.

SentiStrength is primarily based upon a list of words known to normally be used in a positive or negative context. Each word in SentiStrength's lexicon is associated with a

positive or negative score for the polarity and strength of the sentiment term. The score is on a scale of 1 (neutral) to 5 (very positive) or -1 (neutral) to -5 (very negative). For example, love scores +3 and hate scores -4 in this scale. When fed with a new text, SentiStrength checks it for the presence of sentiment terms from its lexicon and predicts the sentiment of the text based upon the scores of the words found, subject to about 12 additional rules (Thelwall, Buckley, & Paltoglou, 2012). Each text is given two scores, one for the strength of positivity contained within it (on the 1 to 5 scale) and one for the strength of negativity (on the -1 to -5 scale). Hence the sentence "I love to hate you" would score -4 and +3.

The results of a sentiment analysis program applied to a Twitter corpus can be used to identify trends in sentiment. This is normally achieved by plotting the average sentiment over time in a time series graph. For example, one study investigated whether peaks of interest in a major media event were associated with increases in positive or negative sentiment (Thelwall, Buckley, & Paltoglou, 2011). Alternatively, the average sentiment before a particular event could be compared to the average sentiment after it, to assess the impact of the event on tweeting, or the average sentiment of tweets mentioning one keyword could be compared to the average sentiment of tweets mentioning a different one. As illustrated by these examples, the goals of a sentiment analysis typically need to involve comparing the average sentiment of multiple sets of tweets or over time.

As described at <http://mozdeh.wlv.ac.uk>, SentiStrength can be used to conduct a sentiment analysis of a Twitter corpus in two ways. It can be applied directly to the corpus, as saved by Webometric Analyst, to record positive and negative sentiment strengths for each tweet. This would help to compare the average sentiment of one or more sets of tweets – for example the average sentiment strengths could be worked out via a spreadsheet after loading the corpus and sentiments scores into it. Alternatively, SentiStrength can be applied to the tweets indexed by Mozdeh and then Mozdeh will produce a graph of the average positive and negative sentiment strengths of all tweets over time or of all tweets matching a particular query over time, as shown in Figure 2 for the 2011 UK riots.

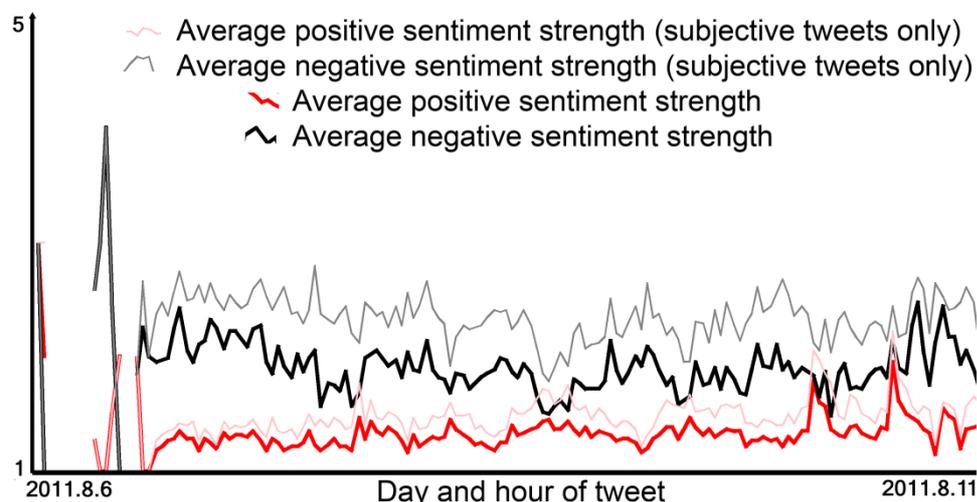


Figure 3. Average positive and negative sentiment strength for UK riots tweets mentioning the police. The tweets are consistently more negative than positive (the negative sentiment strength lines are higher than the positive sentiment strength lines). Negativity seems to decrease over time and there seem to be two important spikes in negativity and two different spikes in positivity towards the end.

Summary

The above description covers how to collect tweets relevant to a topic and how to conduct a sentiment analysis or time series analysis of tweets for that topic. It is designed to give practical advice about what kinds of problems can be investigated and how to conduct analyses using University of Wolverhampton software. Whilst alternative approaches are available, and some are discussed in this book, the methods in this chapter provide an integrated and coherent set that have been used in previously published research. The methods here are quantitative and work best when a reasonably large collection of tweets can be collected for a topic, such as over 10,000. This restriction means that the methods are suitable for topics that are not of purely niche interest and hence generate a reasonable volume of tweeting. When such a corpus can be collected then the methods described here can give insights into how public reactions to the topic change over time, including the identification of significant events, overall trends in topic interest and changes in average sentiment strength over time. Since Twitter can contain spam and a corpus may accidentally contain irrelevant tweets, it is important to check the validity of the results at each stage, for example by examining a proportion of tweets to check for anomalies and spam. With such safeguards, these studies can either give insights into how Twitter is used to discuss an event or insights into the offline event, as mediated through tweeting.

References

- Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In N. Nicolov, & J. G. Shanahan (Eds.), *Proceedings of the 5th international conference on weblogs and social media (ICWSM 2011)* (pp. Retrieved December 6, 2011 from: <http://sm.rutgers.edu/pubs/becker35-icwsm2011.pdf>). Menlo Park, CA: The AAAI Press.
- Bollen, J., Pepe, A., & Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. *ICWSM11, Barcelona, Spain*, Retrieved June 2, 2011 from: <http://arxiv.org/abs/0911.1583>.
- Chung, J. E., & Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? In W. Burgard, & D. Roth (Eds.), *Proceedings of the twenty-fifth AAAI conference on artificial intelligence (AAAI 2011)* (pp. 1768-1769). Menlo Park, CA: AAAI Press.
- Heverin, T., & Zach, L. (2012). Use of microblogging for collective sense-making during violent crises: A study of three campus shootings. *Journal of the American Society for Information Science and Technology*, 63(1), 34-47.
- Neuendorf, K. (2002). *The content analysis guidebook*. London: Sage.
- Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review*, 31(3), 277-289.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2), 406-418.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.
- Wilkinson, D., & Thelwall, M. (2012). Trending twitter topics in english: An international comparison. *Journal of the American Society for Information Science and Technology*, 63(8), 1631-1646.
- Wilkinson, D., & Thelwall, M. (2011). Researching personal information on the public web: Methods and ethics. *Social Science Computer Review*, 29(4), 387-401.