Twitter, Gambling and Time Sensitive Information

Darryl Woodford
Queensland University of Technology
Australia
dp.woodford@qut.edu.au

Abstract

One suggested use of 'Big Data' has been prediction markets; whether that be predicting the stock exchange, health outbreaks, or box office revenues. This paper considers another prediction market; WTA Tennis, and considers how the use of Twitter by professional sportswomen can provide information to gamblers and gambling operators, enabling them to take advantage of the information before it becomes widely known. To do so, it is necessary to consider how incoming tweets can be profiled to ascertain firstly whether they contain information which may impact upon a current or future prediction market, and secondly the reliability of information contained within them. Through such profiling, relevant tweets can be made available to decision makers more quickly than current manually operated services which provide information for the gambling industry.

Keywords

Twitter; Big Data; Analysis; Gambling

Twitter & Prediction Markets

Twitter has been used as a tool to attempt to predict a range of activities, including heath outbreaks (Ritterman et al., 2009; Polgreen et al., 2007) , Box office revenues (Asur & Huberman, 2010), the Stock Exchange (Zhang et al., 2010), and elections (Tumasjan et al., 2011) . Whilst predictions of health could be considered for the public good, there are also of course revenue considerations, not least for health insurers and those companies producing potential vaccines. In other cases, there is not even an attempt at altruistic purposes, with the sole intention being to out-perform traditional market indicators, in order to return either direct or indirect revenue for those processing the data, or their employers.

Based on standard Twitter analysis methodologies (Bruns & Burgess, 2011, 2012) I analyzed 217,560 tweets that matched the 'WTA' keyword between April and July 2012, with a particular focus on tweets discussing injuries or form. I compared the content of these tweets, their timing, and their distribution by others (e.g. retweets by journalists) to movement in betting markets for future matches and tournaments. In many cases, the timing of the tweet preceded a significant change in the prediction market. I also considered the role of 'amplifiers'; television or internet sports personalities or writers who, as well as breaking their own news, re-tweet that of colleagues.

Value of Information in the Gambling Industry

There are clear advantages to the use of twitter by gambling industry participants to gather information on player injuries, team selection and other factors. While in the US context much of this is performed under the guise of fantasy sports information (which is league sanctioned), in other markets the significance of information for gambling is acknowledged, both by those distributing the information and through platforms which aggregate it, such as *TennisForm.com*.

The value of 'inside information' is clear. While NBA referee Tim Donaghy received prominent attention for potentially altering the outcome of matches he was in charge of, his account also acknowledges (Donaghy, p. 4) he had an "inside advantage because of [his] access to pregame meetings. It was common for my fellow referees to voice their opinions about who they expected to win on a given night. Those opinions were often based on their knowledge of confidential inside information pertaining to players and teams, such as injury reports unknown to the general public”.

Many professional sporting leagues, including all of the major United States sports and Australia’s NRL and AFL have strict reporting requirements for injuries. Whilst the NFL, in line with their anti-gambling stance, asserts in their bylaws that the purpose of injury reports are to “tell the opposing coach of the injury status of his players so that each coach can plan strategies for the game” (Harding, 2006), the use of such information in gambling and fantasy sports is undeniable. Borghesi et al. (2009), in a study considering Arena Football, noted how player injury status in the sport meant that “diligent bettors may, at times, have an information advantage over bookmakers”. The significance of this information for predicting sporting events is then clear.

Market Adjustments

Whilst mainstream betting sports such as the NFL, NBA, College Football, College Basketball and Soccer would have their markets impacted by injury developments, this current study focuses on tennis, a sport where the entire outcome of a match, and a significant impact on tournament markets, is dependent on a single player. A large number of players on both the ATP and WTA tour have twitter accounts, and these accounts are frequently used to discuss niggling injuries, upset stomachs, flu symptoms, or to simply announce a withdrawal from an upcoming tournament. Tweets such as “"Mica has recovered. We have trained together. I believe everything is OK." (Nenad Zimonjic about Michael Llodra, 19 March 2012), or “Update: feeling back to normal and practicing” (Vania King, 17 March 2012) will inevitably have an impact on betting markets while not being stories which would attract mainstream press coverage.

Additionally, the range of markets currently offered, particularly online, means that any tweets of this kind would have an impact on prediction markets, whether it be the market for a match the same day, the outright market for an upcoming tournament, or a long term performance market, such as the odds for a player to win a tournament in the current calendar year.

Whilst sites such as *Tennis Form* aggregate data on player injuries (at a cost of 50 Euros per month), the ability to filter and access this information in real time is key to obtaining an advantage on the remainder of the market. In this way, the same question is raised as with the stock exchange, with health, or with crisis management; how do you filter the information and place the result in front of somebody who can decide what to do with it.

Selecting Twitter Information

Whilst analysis of past events is useful, it is more relevant to consider how such tweets could be isolated for future use. While the volume of tweets directly from WTA approved player accounts is relatively low, at between 50 and 200 tweets per day, the broader conversation (which includes coaches, agents, training partners etc) on the WTA keyword amounted to a further 2000/day, and this would rise significantly if other keywords or accounts were added to the data set. Within these tweets is useful information, but also substantial rumor and misinformation. Indeed, the network graph (Figure 1) shows a large number of players, but also news organizations and fans prominent in the conversation, whose guesses at injury status may not be reliable.

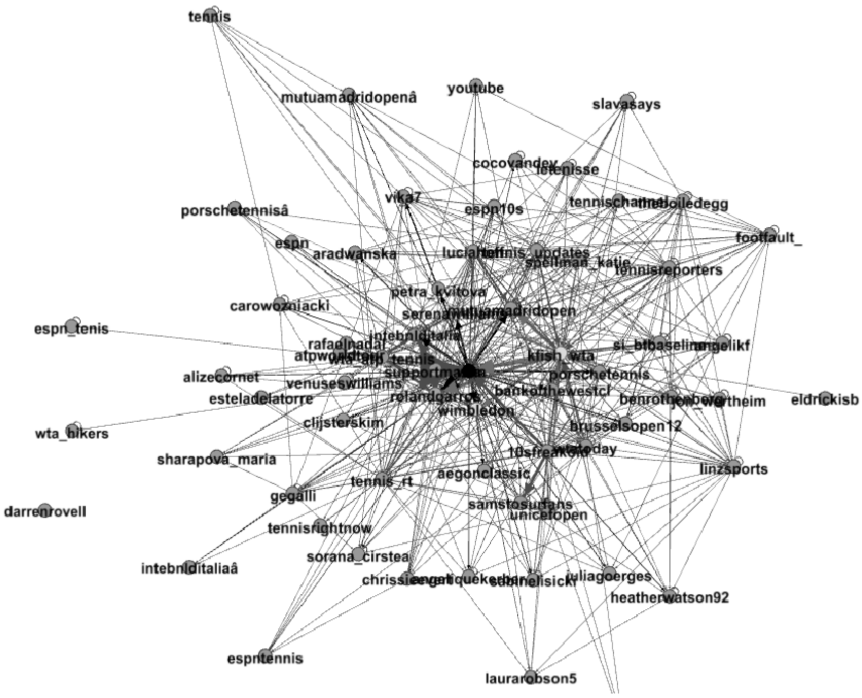

Figure 1: Twitter Network -- #WTA hashtag, April-July 2012

However, through an analysis of past tweets, there appear to be key indicators as to the reliability of information, which, whilst requiring manual judgment to set up, may enable automated filtering. Those tweets directly from the players account can be generally assumed reliable (unless the player has a history of misinformation). A mapping of the players immediate social and professional network (who have twitter accounts) can be produced through an analysis of their @ replies on Twitter, and thus a reliability score can also be assigned to these accounts, in proportion to the volume of tweets between them and the player. Finally, external sources such as travelling reporters and photographers, as well as event staff for each WTA tournament must be considered; a process which involves both research and qualitative judgment. In all cases, a retweet by one of the 'significant' accounts has value; if the retweet is of another trusted source this would serve to amplify the information, whilst a retweet of a third party would raise attention, but require further validation.

Conclusion

Bookmakers and gamblers alike are charged with evaluating each piece of information, the source, and the potential impact on the betting market; a service with a value, and a significant potential for profit. The stakeholders, their requirements and a detailed understanding of the domain is necessary to consider how developments in automatically ranking tweets could be applied to any given set of information, whether a betting market, health scare or crisis situation.

References

Asur, S., & Huberman, B. (2010). Predicting the Future with Social Media. In *Proceedings of the ACM Conference on Web Intelligence, 2010.* Retrieved from http://arxiv.org/pdf/1003.5699.pdf.

Borghesi, R., Paul, R., & Weinbach, A.P. (2009). Market Frictions and Overpriced Favorites: Evidence from Arena Football. In *Applied Economics Letters,* 16(9): 903-906

Bruns, A. & Burgess, J. (2011). Tools. Retrieved from http://mappingonlinepublics.net/resources/.

Bruns, A. & Burgess, J. (2012). Researching News Discussion on Twitter: New Methodologies. Journalism Studies, 13.5-6.

Donaghy, T. (2010). Personal Foul: A First-Person Account of the Scandal that Rocked the NBA. Sarasota, FL: Four Daughters LLC.

Harding, Casey N. (2006). *Nickel and Dimed: North Carolina court blocks Carolina Panthers' attempt to avoid payment of workers' compensation benefits to injured athletes.* In 28 N.C. Cent. L.J. 241

Polgreen, P.M., Nelson, F.D., Neumann, G.R., & Weinstein, R.A. (2007). Use of Prediction Markets to Forecast Infectious Disease Activity. In *Clinical Infectious Diseases,* 44(2):272-279.

Ritterman, J., Osborne, M. & Klein, E. (2009). Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic. In *Workshop on Mining Social Media*. Retrieved from http://www.christopia.net/data/school/2011/Fall/social-media-mining/project_proposal/sources/ritterman-2009.pdf

Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2011). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. In *Social Science Computer Review*, 29(4):402-418.

Zhang, X., Fuehres, H., & Gloor, P.A. (2010). Predicting Stock Market Indicators Through Twitter- I Hope it is not as bad as I fear. In *Collaborative Innovations Networks Conference (COINs)*.

From "Big" to Reasonable Data: Collecting and Extracting Data From An Archive of a Social Movement

Shawn Walker
University of Washington
United States
stw3@uw.edu

Abstract

Collecting Twitter data and following changes in an ongoing, dynamic social movement, such as the Occupy Wall Street movement, is a complex task. It involves the development of technical infrastructure to collect and make the tweets available for exploration and analysis. A strategy to respond to changes in the social movement is also required or the resulting tweets will only reflect the discussions and strategies the movement used at the time the keyword list is created — in a way, keyword creation is part strategy and part art. In this paper we describe strategies for the creation of a social media archive, specifically tweets related to the Occupy Wall Street movement, and methods for continuing to adapt data collection strategies as the movement's presence in Twitter changes over time. We also discuss the opportunities and methods to extract smaller slices of data from an archive of social media data to support a multitude of research projects in multiple fields of study. These methods include the use of timeslices, keywords, hashtags, URL coding, metadata matching, geolocation, geocoding, and geomapping.

Keywords

twitter, social movements; big data; social media; archiving

Introduction

On June 2nd, 2011, Adbusters proposed a peaceful demonstration, "Occupy Wall Street", to take place on September 17th to demand a separation of money from politics. Over the course of the next three months, face-to-face working groups met in NYC to create a General Assembly to coordinate and organize action. Facebook pages for Occupy city camps sprung up in early October, accumulating tens of thousands of likes in major cities such as Philadelphia and Chicago (Caren & Gaby, 2012). Twitter accounts and hashtags such as #occupywallst, #ows, #occupy, and #needsoftheoccupiers emerged to facilitate the coordination and exchange of relevant information and requests. While on-the-ground efforts and participation were critical to sustaining and growing the movement; digital tools and technologies acted as communication infrastructure, providing channels to share information, organize events, coordinate activity, and connecting participants and camps to one another (Author et al, 2012).

Collecting Twitter data and following changes in an ongoing, dynamic social movement, such as the Occupy Wall Street movement, is a complex task. It involves the development of technical infrastructure to collect and make the tweets available for exploration and analysis. A strategy to respond to changes in the social movement is also required or the resulting tweets will only reflect the discussions and strategies the movement used at the time the keyword list is created — in a way, keyword creation is part strategy and part art. In this paper we describe strategies for the creation of a social media archive, specifically tweets related to the Occupy Wall Street movement, and methods for continuing to adapt data collection strategies as the movement's presence in Twitter changes over time. Since most studies focus on specific aspects of a social movement, it is not possible or desirable to analyze all of the tweets within an archive. We also discuss opportunities and methods to extract smaller slices of data from an archive of social media data to support a multitude of research projects in multiple fields of study.

The Occupy Movement's Presence in Social Media

Occupy used many different online and offline channels to communicate, request resources, and coordinate between city sites and with the public at large. These included email and mailing lists, public general assemblies meetings at individual camps, Twitter accounts, Facebook pages, blogs, websites, and phone trees. Some of these methods of communication, such as Twitter and Facebook, leave artifacts that can be studied after an event, while other methods of communication are more ephemeral requiring researchers to observe interactions as they take place. While the media, in many cases, treated the Occupy Wall Street movement as a one monolithic movement; in actuality it consisted of separate camps in many different cities. Each camp used the communication channels best suited its needs; with its strategy evolving over time. For example in the case of Seattle, their twitter account was active at the start of their occupation, switching to their blog, and then finally to their Facebook page. After they switched to Facebook, their twitter presence atrophied. A multitude of hashtags were used for each city — most cities had a hashtag, a Twitter account, and also used national and international hashtags such as @ows. This diagram below (Figure 1) social media, and Twitter in particular, were centers of communication for the Ouccpy movement.

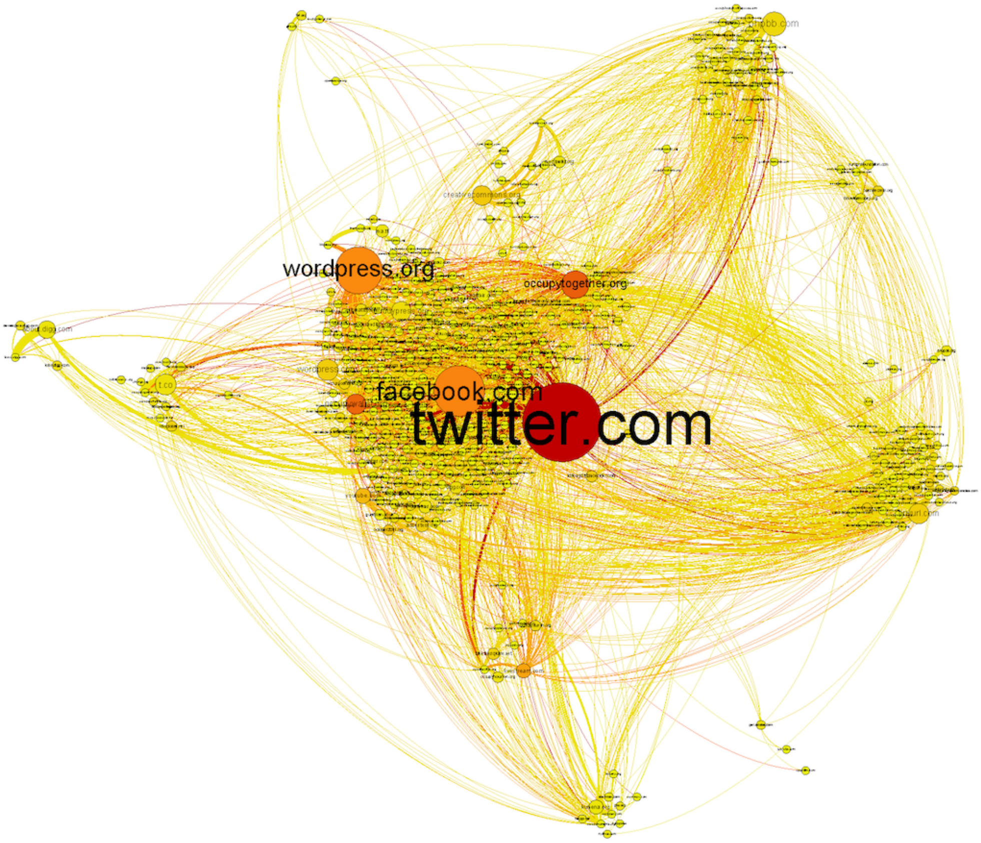

Figure 1: Network diagram of the links between Occupy Wall Street Movement related websites. Nodes represent websites, edges represent links between sites.

Data Collection and Keyword Curation

Twitter data was collected from mid October, 2011, when a national network of occupations had been established and protest activity was at its peak, through the end of 2012 when the national protests had tailed off into more scattered local actions. The peak activity period captured in our data from October 19, 2011, to December 31, 2011 contained roughly 20 million Tweets (20,645,921 to be exact). The archive collected tweets using Twitter's Streaming API track method, which returns tweets matching any of the search keywords occurring in the text, hashtags, @mentions, or URLs within a tweet.

A curated list of 102 hashtags, keywords, and Occupy city accounts related to the Occupy movement was created and updated by a panel of faculty and graduate students involved in this project. The initial keyword list was created by conducting preliminary fieldwork of Occupy's existing online presence in Twitter and social media (Author, 2013). Popular hashtags and keywords were seeded after a review by the panel of researchers involved in the project. The resulting data stream was examined at regular intervals for emerging hashtags and keywords. New terms were added, but no keywords were removed from the list, after being reviewed by the entire research team, resulting in dynamic archive based on a list of 355 keywords as the collection continued through the end of 2012, by which time the database included more than 75 million tweets.

A script maintained a constant connection to the Twitter Streaming API; writing the tweets received to a file. At the end of each day, a new data file was created and the file from the previous day was backed-up and processed. The metadata added to each tweet in the processing step included:

- Expansion of shortened URLs
- List of hashtags in the tweet
- List of mentions in the tweet
- Count of the number of hashtags, URLs, and mentions in the tweet
- List of collection keywords matching the tweet text.

Discussion

After data was inserted into the archive, researchers across multiple disciplines including Information Science, Communication, Geography, and Statistics expressed an interest in the social media presence of the Occupy movement. As each team presented their research questions, a plan for data extraction was created. While some teams used specific hashtags and time slices, other teams used URL coding, metadata matching, geolocation, geocoding, and geomapping to slice data form the archive. Some teams used combinations of these options to produce very specific and narrowly defined datasets; narrowing down from 70+ million tweets to hundreds or thousands of tweets.

The apparent ease with which tweets may be aggregated belies the difficulty of designing a reliable, reproducible data collection strategy. These characteristics are then linked to broader issues of designing research for big social data, and the emerging "digital divide" in access to data (Manovich, 2011; boyd & Crawford, 2011). Conversely, long-term observation remains one of the areas in which this field has been the weakest (boyd & Crawford, 2011, 4). The creation of social media archives and broader research collaborations can start to address the "digital divide" in access to data, technical expertise, allow for longitudinal data collection, and supports the exploration and extraction of smaller data sets using a wide variety of extraction methods which are more appropriate to the research questions at hand. As a result, groups of researchers across multiple fields are able to benefit and experimentation with the larger dataset, but use different slices of the same dataset more appropriate to their object, methods, and tools of study.

References

boyd, d. & Crawford, K. (2011). Six Provocations for Big Data. Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" on September 21, 2011.

Bruns, A (2007) Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool. *First Monday*, 12(5).

Caren, N. & Gaby, S. Occupy Online: Facebook and the Spread of Occupy Wall Street (October 24, 2011). Available at SSRN: http://ssrn.com/abstract=1943168 or http://dx.doi.org/10.2139/ssrn.1943168.

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the Digital Humanities*. Minneapolis, MN: The University of Minnesota Press.

Author. (2012).

Author. (2013).

Extracting important information from a social network stream during crisis

Avijit Paul
Queensland University of Technology
Australia
a1.paul@qut.edu.au

Abstract

The first paper considers possible coding mechanism for incoming tweets during a crisis, taking a large stream of incoming tweets and selecting which of those needs to be immediately placed in front of responders, for manual filtering and possible action. The paper suggests two solutions for this, content analysis and user profiling. In the former case, aspects of the tweet are assigned a score to assess its likely relationship to the topic at hand, and the urgency of the information, whilst the latter attempts to identify those users who are either serving as amplifiers of information or are known as an authoritative source. Through these techniques, the information contained in a large dataset could be filtered down to match the expected capacity of emergency responders.

Keywords

Natural disaster; crisis; twitter; content analysis; user profiling

Introduction

When an unexpected crisis such as earthquake happens, the amount of information in Twitter becomes extraordinarily large (Figure 1) in a very short period of time (Hendrickson, 2012). This large volume of data creates an opportunity but also brings the problem of extracting important information. This panel paper suggests a framework that can be used to identify important information from these massive datasets in times of crisis.

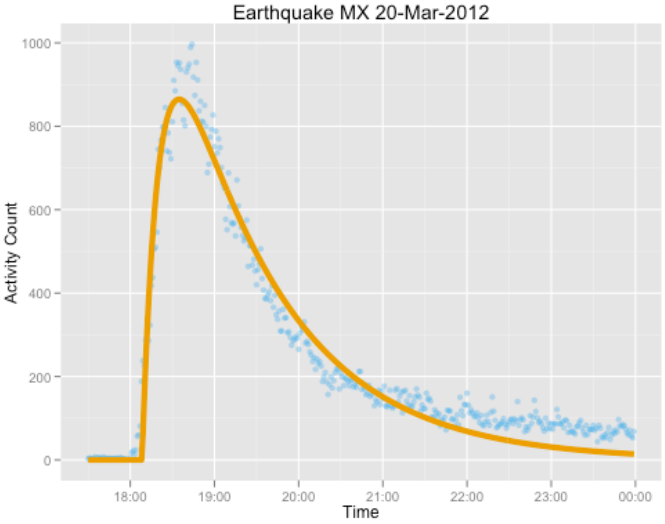

Figure 1: People's response for Mexico Earthquake, as a sudden event (Hendrickson, 2012).

Finding Patterns

There are two approaches commonly used to find patterns in crisis datasets generated via Twitter. The first is content analysis, and the second is user profiling.

The content analysis approach also has two aspects to it. The first is the manual approach; based on how, in different case studies (Bruns, Burgess, Crawford, & Shaw, 2012; Vieweg, 2012), tweets could be grouped in broad categories such as information, awareness, advice, environment status etc. Although this provides a good understanding of the type of tweets generated in crisis situations, it is almost impossible to replicate with a large volume in a short period of time.

The second approach in content analysis is the automated approach where real time data processing methods such as dictionary-based, rule based, and hybrid methods have been used to find patterns or named entities (location, person or organization) (Döhling & Leser, 2011; Song, Tjondronegoro, & Docherty, 2012). This method has its own limitations, as the meaning and the context rapidly changes in Twitter (Vlachos, 2011).

User profiling looks at the users who are contributing to the dataset. As the authoritative users (or actors) are highly central and active, they tend to transmit the information very quickly. Therefore current work (Cha, Haddadi, Benevenuto, & Gummadi, 2010; Garcia-Herranz, Egido, Cebrian, Christakis, & Fowler, 2012; Wagner, Liao, Pirolli, Nelson, & Strohmaier, 2012; Yamaguchi, Takahashi, Amagasa, & Kitagawa, 2010) suggests that identifying, profiling and following authoritative users is superior to crunching enormously large datasets to identify what is important. An additional benefit of user profiling is that it is very expensive to purchase access to Firehose (100% of Twitter data) or even Decahose (~10% of Twitter data) to identify potential warning signs or important keywords (Reips & Garaizar, 2011).

Methodology

Combining both of these approaches is useful in identifying which messages are important at the time of crisis. In this paper, I identify a weighting criterion for every incoming tweet that is captured. By assigning weights to each variable the proposed framework identifies which combination yields the most important results for a crisis situation. If the tweet gains more than a certain weight, it is deemed important and is sent for further evaluation by a person in charge.

Keyword from Volume

Below (Table 1) is a sample weight assignment to identify if a tweet should be marked as a "crisis tweet" for further processing or it should be ignored.

Table 1: Weight assignment to an incoming tweet

No	Category	Reference	Weight
1	keyword	Contains disaster related keyword such as "earthquake"	+ 10
2	#hashtag	Includes a hashtag	+ 15
3	#Disaster-hashtag	Hashtag contains a disaster related word	+ 8

Therefore if an incoming tweet contains any of the keywords or hashtag it can be sent to the individual tweet analysis to gain a better understanding of whether this is a crisis situation or if it is a false alarm.

Categorising Individual Tweets

Once a tweet is sent for detailed analysis, it can then be given further weight in order to gain a better understanding of what this tweet actually contributes. For example, a tweet sent by the "keyword from volume" separator described above can now be analyzed in detail (Table 2):

Table 2: Weight assignment to an individual tweet

No	Category	Reference	Weight
1	RT	Retweet of a previously stored tweet	-10
2	Named entity	Name of a place not identified before	+10
3	URL	Uses a link to a news site	+ 2
4	Instragram URL	Is that a cat photo?	- 5
5	Vine video	A new vine video just recorded	+ 3
6	Temporal word	Has word with certain time	+ 1
7	Appeal word	Included word “help”	+10

Based on these criteria, if it appears to be a potential crisis related tweet it then gets a high total mark and is displayed on the dashboard, while at the same time the identified keyword or hashtag is added to the data collction stream to understand if it is a rising or a falling trend.

Identifying if a crisis is on the rise

Once a keyword, named identity or hashtag from the individual tweet analysis has been identified as potentially a crisis related tweet, it is sent to this section to identify the incoming rate of this tweet. Based on how frequently it has been mentioned it is then given a new weight (Table 3).

Table 3: Weight assignment to an incoming tweet based on the occurance

No	Category	Reference	Weight
1	Double mention in last 5 seconds	Potentially dangerous	+ 10
2	200 mention in last one minute	May be a crisis is arising	+ 25
3	Mention + RT from news organization	Information already known	- 10

Therefore, based on the frequency of the mention, it can now be identified whether this is a trending tweet or it is actually declining in the stream.

User Profling

Once a tweet has been identified as a seemingly potential crisis tweet via keyword or hashtag, user profiling is activated. This is to identify the authority, or the role, the user might have within the situation. Based on the weights (Table 4), they can be categorised into leaders, novices or insiders for that situation. If they appear to be an authority, their tweet gets higher priority than someone who is a novice.

Table 4: Weight assignment to a user

No	Category	Reference	Weight
1	New registration	Spam bot or real human	- 2
2	User has above 500 followers	Popular user	+ 5

3	User has location enabled	The location matches the named entity mentioned in the tweet	+ 10
4	User bio has listed keywords (such as emergency, research etc)	Potentially an expert user	+ 15
5	User is included in other user's curated list	Known expert	+15

In the end, if the total weight is above an identified level, it is identified as important and it can now be sent to the crisis monitoring authority for evaluation.

Discussions on findings

The significance of this framework is the ability to separate tweets that are potentially important for the target audience (community safety, red cross etc) from the stream of a noisy twitter dataset. This approach is currently being tested with archived datasets and the full results will be published when available. Based on preliminary results, the combination of "keyword, named entity and URL with images that do not contain Retweets" appears to be more important that a combination of "hashtag, Retweet and temporal word". User profiling still remains challenging due to the difference in activity among users from the archived dataset at the current time. However, as this is ongoing research, further results are expected in the coming months.

References

Bruns., A., Burgess, J., Crawford, K., & Shaw, F. (2012). CCI Floodsreport. Retrieved from

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010). Measuring user influence in twitter: The million follower fallacy. In 4th international aaai conference on weblogs and social media (icwsm) (Vol. 14, pp. 8).

Döhling, L., & Leser, U. (2011). EquatorNLP: Pattern-based Information Extraction for Disaster Response.

Garcia-Herranz, M., Egido, E. M., Cebrian, M., Christakis, N. A., & Fowler, J. H. (2012). Using Friends as Sensors to Detect Global-Scale Contagious Outbreaks. arXiv:1211.6512. Retrieved from http://arXiv.org/abs/1211.6512.

Hendrickson, S. (2012). Gnip The Social Cocktail, Part 2 Expected vs. Unexpected Events. Retrieved from http://blog.gnip.com/expected-vs-unexpected-events-in-social-media/

Reips, U.-D., & Garaizar, P. (2011). Mining twitter: A source for psychological wisdom of the crowds. Behavior research methods, 43(3), 635-642.

Song, W., Tjondronegoro, D. W., & Docherty, M. (2012). Understanding user experience of mobile video: framework, measurement, and optimization. Mobile Multimedia: User and Technology Perspectives, 3-30.

Vieweg, S. (2012). Twitter communications in mass emergency: contributions to situational awareness. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion (pp. 227-230): ACM.

Vlachos, A. (2011). Evaluating unsupervised learning for natural language processing tasks. Paper presented at Proceedings of the First Workshop on Unsupervised Learning in NLP, Edinburgh, Scotland.

Wagner, C., Liao, V., Pirolli, P., Nelson, L., & Strohmaier, M. (2012, 3-5 Sept. 2012). It's Not in Their Tweets: Modeling Topical Expertise of Twitter Users. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom) (pp. 91-100).

Yamaguchi, Y., Takahashi, T., Amagasa, T., & Kitagawa, H. (2010). Turank: Twitter user ranking based on user-tweet graph analysis. Web Information Systems Engineering–WISE 2010, 240-253.