Real-time topic detection with bursty n-grams: RGU's submission to the 2014 SNOW challenge.

MARTIN, C. and GOKER, A.

2014

Real-time topic detection with bursty n-grams: RGU's submission to the 2014 SNOW Challenge

Carlos Martin, Ayse Göker
IDEAS Research Institute
School of Computing & Digital Media
Robert Gordon University, Aberdeen AB10 7QB
{c.j.martin-dancausa, a.s.goker}@rgu.ac.uk

Abstract

Twitter is becoming an ever more popular platform for discovering and sharing information about current events, both personal and global. The scale and diversity of messages makes the discovery and analysis of breaking news very challenging. Nonetheless, journalists and other news consumers are increasingly relying on tools to help them make sense of Twitter. Here, we describe a fully-automated system capable of detecting trends related to breaking news in real-time. It identifies words or phrases that 'burst' with sudden increased frequencies, and groups these into topics. It identifies a diverse set of recent tweets that are related to these topics, and uses these to create a suitable human-readable headline. In addition, images coming from the diverse tweets are also added to the topic. Our system was evaluated using 24 hours of tweets as part of the Social News On the Web (SNOW) 2014 data challenge.

1 Introduction

The growth of social networking sites, such as Twitter, Facebook and Reddit, is well documented. Every day, a huge variety of information on different topics is shared by many people. Given the real-time, global nature of these sites, they are used by many people as a primary source of news content [New11]. Increasingly, such sites are also used by journalists, partly to find and track breaking news but also to find user-generated content such as photos and videos, to enhance their stories. These often come from eyewitnesses who would be otherwise difficult to find, especially given the volume of content being shared.

The Social News On the Web (SNOW) 2014 task is to provide journalists with headlines and supporting information from Twitter to summarize and highlight news stories from Twitter in near real-time. This paper describes the methods we used and provides some initial evaluation. However, the official evaluation results of our approach in the challenge can be found in [PCA14]. We describe a practical tool to help journalists (and other news readers) to find newsworthy topics from message streams without being overwhelmed. Note that it is not our intention to re-create Twitter's own "trending topics" functionality, that is usually dominated by very high-level topics and memes, defined by just one or two words or a name and with no emphasis on 'news'.

Our system works by identifying words or phrases that show a sudden increase in frequency (a "burst") and then finding co-occurring groups to identify topics. Such bursts are typically responses to real-world events. In this way, the news consumer can avoid being overwhelmed by redundant messages, even if the initial stream is formed by diverse messages. The emphasis is on the temporal nature of message streams as we bring to the surface groups of messages that contain suddenly-popular phrases. An early version of this approach was recently described [APM+13, MCG13], where it compared favourably to several alternatives and benchmarks including LDA (Latent Dirichlet allocation). We have also demonstrated our approach by finding events in football matches, both 'objective' event detection [CMG14a] and from the different perspectives of each team's fans [CMG14b]. Here we in-

In: S. Papadopoulos, D. Corney, L. Aiello (eds.): Proceedings of the SNOW 2014 Data Challenge, Seoul, Korea, 08-04-2014, published at http://ceur-ws.org

clude improvements to the topic detection approach, the topic labelleling (i.e. adding human-readable headlines to stories) and the display of diverse, relevant tweets and images within each topic.

2 Related Work

Social media are now central to the work of news professionals, such as tracking stories on Twitter or Facebook, and hosting live blogs of ongoing events [New10] . Newman also describes the growth of collaborative, networked journalism, where news professionals draw together a wide range of images, videos and text from social networks and provide a curation service. Broadcasters and newspapers can also use social media to increase brand loyalty across a fragmented media marketplace.

Petrovic et al. [POL10] focus on the task of first-story detection (FSD) or "new event detection". They use a locality sensitive hashing technique on 160 million Twitter posts, hashing incoming tweet vectors into buckets in order to find the nearest neighbours and so detect new events and track them. This work is extended in Petrovic et al. [POL12] using paraphrases for first story detection on 50 million tweets. Their FSD evaluation used newswire sources rather than Tweets, based on the existing TDT5 datasets. The Twitter-based evaluation was limited to calculating the average precision of their system, by getting two human annotators to label the output as being about an event or not. This contrasts with our goal here, where our results will be compared against a fixed set of topics to estimate the topic-level recall, i.e. to count how many newsworthy stories the system retrieved.

Benhardus [Ben10] uses standard collection statistics such as *tf-idf*, unigrams and bigrams to detect trending topics. Two data collections are used, one from the Twitter API and the second being the Edinburgh Twitter corpus containing 97 million tweets, which was used as a baseline with some natural language processing used (e.g. detecting prepositions or conjunctions). The research focused on general trending topics (typically finding personalities and for new hashtags) rather than focusing the needs of journalistic users and news readers.

Shamma et al. [SKC11] focus on "peaky topics" (topics that show highly localized, momentary interest) by using unigrams only. The focus of the method is to obtain peak terms for a given time slot when compared to the whole corpus rather than over a given time-frame. The use of the whole corpus favours batch-mode processing and is less suitable for real-time and user-centred analysis, whereas our work here is designed for use in a live, real-time system.

Becker et al. [BNG11] also consider temporal issues by focusing on the online detection of real world events, distinguishing them from non-events (e.g. conversations between posters). Clustering and classification algorithms are used to achieve this. Methods such as n-grams and NLP are not considered.

Zhang et al. [ZXM+] consider threads of tweets forming conversations. They concentrate on cleaning and merging topics to filter out a thread, and merging these to create global topics; replies and follow-up postings are used as evidence to assist this process. They collected data from the Sina Weibo microblog, with 1,100 threads in 16.5k postings.

3 Methods

In this section we describe various aspects of our proposed approach to topic detection and discuss how they work together. We consider "temporal document frequency-inverse document frequency" as a variation of the classic *tf-idf* to find trending terms at a specific point in time. We discuss methods to group these terms into topic-specific clusters and the use of n-grams to find phrases rather than isolated terms. We then describe adding human-readable labels to the topics, assigning diverse but relevant tweets to each topic and ranking the topics by significance, to avoid overwhelming the user.

3.1 BNgrams

Term frequency-inverse document frequency, or *tf-idf*, has been used for indexing documents since it was first introduced [SJ72]. We are not interested in indexing documents however, but in finding novel trends, so we want to find terms that appear more often in one time period than in others. We treat temporal windows (i.e. the set of all tweets posted between a start and end time) as documents and use them to detect words and phrases that are both new and significant. We therefore define newsworthiness as the combination of novelty and significance. We can maximise *significance* by filtering tweets either by keywords or/and by following a carefully chosen list of users, and maximise *novelty* by finding bursts of suddenly high-frequency words and phrases. This approach makes more sense in Twitter space due to number of characters limitation (140) making messages more focused and the high number of retweets and post copies of previous tweets.

This approach indexes all keywords from the posts of the collection, apart from other metadata, such as hashtags, entities, urls... The keyword indices, implemented using Solr[1], are organized into different time slots. In this approach, the index considers bigrams and trigrams for the post text. Once, the index is

[1] https://lucene.apache.org/solr/

created, we select terms with a high "temporal document frequency-inverse document frequency", or *df-idf*$_t$. The *df-idf*$_t$ score is computed for each term of the current time slot i based on its document frequency for this time slot and penalized by the logarithm of the average of its document frequencies in the previous s time slots:

$$df{-}idf_{ti} = (df_{ti} + 1) \cdot \frac{1}{\log\left(\frac{\sum_{j=i}^{s} df_{t(i-j)}}{s} + 1\right) + 1}. \quad (1)$$

where s=2 after doing some preliminary experimentation (where s=1, 2, 3 were tested). This produces a list of terms which can be ranked by their *df-idf*$_t$ scores. Note that we add one to term counts to avoid problems with dividing by zero or taking the log of zero. Based on experiments reported previously [APM+13, MCGM13] and subsequent work in the last months, we use entities, hashtags, urls, and n-grams as terms in this work. To maintain some word order information, we consider n-grams, i.e. sequences of n words. 2- and 3-grams are better options according to previous results [MCG13]. In general terms, we consider those terms that can be useful to describe a story properly. Regarding the entities, our approach includes Stanford 3-class named entity recognizer [FGM05] to detect person, organization and location entities. One of the strong points of using this algorithm is its efficiency as our target is to implement a system working in real-time. High frequency terms are likely to represent semantically coherent phrases. Having found bursts of potentially newsworthy terms, we then group together terms that tend to appear in the same tweets. Each of these clusters defines a topic as a list of terms.

3.2 Topic Clustering

An isolated word or phrase is often not very informative, but a group of them can define the essence of a story. Therefore, we group the most representative terms into clusters, each representing a single topic. A group of messages that discuss the same topic will tend to contain at least some of the same terms.

Hierarchical clustering, implemented in earlier versions of this approach [APM+13, MCG13] assigns each term to exactly one topic cluster. This may cause problems such as if one term (e.g. "Obama") is part of more than one simultaneous story (e.g. "Obama wins in Ohio" and "Obama wins in Illinois"). Several clustering algorithms allow for "partial" membership, such as Apriori algorithm. A preliminary study has been performed considering different topic clustering approaches for our approach [MCGM13].

The Apriori approach [AS+94] finds all the associations between the most representative terms based on the number of tweets in which they co-occur. Each association is a candidate topic at the end of the process. In addition, no specific number of associations has to be specified in advance.

One parameter associated to this technique is *support* value which determines the minimum number of documents a group of terms (association) should share to be considered as an association or candidate topic. The value of this parameter represents a percentage of all the documents from the corresponding slot. Preliminary experiments considering different values of this parameter suggested to fix its value to 0. It means that no association/candidate topic is discarded.

In addition, maximal associations are obtained at the end of the approach to avoid overlaps in the final candidate topics set. We want one news story to appear as one candidate topic, and not to appear in several (near-)duplicate topics at once. The main idea of this approach is to remove all the associations whose keywords are contained in another association and sharing most of the topic tweets ($> 70\%$) with the previous one. This second requirement was introduced to confirm that both topics are talking about the same matter. This exact value is not critical, according to preliminary experiments.

To avoid possible duplications of similar topics in the same timeslot, a topic merging process was implemented to detect similarities between topics, based on the matching of similar topic terms (keywords, entities, hashtags and URLs). The terms are aggregated per topic and the most frequent ones are selected to be used in the merging process. The selection of a "partial" membership clustering approach makes the matching process between two topics easier as several topics could share the same term.

3.3 Topic labeller

The editor of Buzzfeed recently told an audience of media specialists at Harvard University that headlines look increasingly like tweets[2]. Following this advice, we form a human-readable title for each topic by finding a representative tweet and then performing minimal editing for clarity. We score the set of tweets associated with a topic based on the number of topic terms (computed in the previous step) each tweet contains and number of times this tweet is duplicated in the timeslot. The similarity between tweets is computed as cosine similarity. The combined score is a weighted sum of these two components:

[2]http://perryhewitt.com/5-lessons-buzzfeed-harvard/

Raw tweet	*Clean tweet*
:(RT @civicua: Free #Venezuela ! #Ukraine is with you! #euromaidan Photo by Liubov Yeremicheva http://t.co/CRDoME5eOb	Free #Venezuela ! #Ukraine is with you! #euromaidan Photo by Liubov Yeremicheva
Microsoft releases Office 2013 Service Pack 1 http://t.co/9R4JiVSguk by @epro	Microsoft releases Office 2013 Service Pack 1

Table 1: Example of topic label based on clean tweet

Topic	*Useful Replies*
Researchers Are Building a Lie Detector For Twitter	- @mashable no more fake accounts following us hopefully ! - @mashable twitter is a sanctuary for irony.. this is pointless
Topic	*Spam reply*
Nigerian Islamists kill 59 pupils in boarding school attack	@Reuters Man... Kanye was ugly in high school. See the pics http://t.co/od77TGWQYx

Table 2: Examples of the use of replies to populate topics

$$tw.\ score = \alpha \cdot Norm.\ topic\ terms + (1-\alpha) \cdot Norm.\ duplicates \quad (2)$$

where we use $\alpha = 0.8$ to give more importance to the number of topic terms per tweet, as this tweet would be more connected to the main story of the topic. The two factors are normalized by dividing by the number of topic terms and the number of tweets in the timeslot respectively. The tweet with the highest score is selected and its text, after cleaning it (user mention, URL and abbreviations, such as RT and MT, removal), is set to the topic title. Examples of tweets before and after cleaning are shown in Table 1.

3.4 Topic Items Population

As the tweet collection was indexed in Solr, we built Solr queries for each topic, composed of the most representative terms (entities, hashtags, URLs and n-grams) to retrieve all the tweets associated to this topic. In case of topics containing n-grams and entities, we consider those keywords close to the entities in the final Solr query to avoid noisy terms. We limit this selection to tweets whose publication time is earlier than the end time of the corresponding timeslot to simulate a live, real-time scenario. In addition, replies to these tweets are also considered as they can add people's view about the topic that could not be retrieved using text-dependent queries. In some occasions, popular tweets (with many retweets and replies) contain some spam replies with advertising purposes in most cases (see examples in Table 2).

To get diverse tweets per topic, we compute the cosine similarity between each pair of tweets and consider a threshold β to remove duplicates. After some preliminary experiments, we set β to 0.7. In addition, retweets are replaced with the original tweets if they are in the collection (otherwise the retweet is kept as-is). Table 3 shows some examples of tweets with and without this diversity selection process where the last two tweets from the first column are ignored in the second one because they look similar to the first two tweets of the list.

Lastly, images are also retrieved per topic. If diverse tweets contain link/s to image/s in their metadata, they are also added to the topic details.

3.5 Topic Ranking

To maximise usability we need to avoid overwhelming the user with a very large number of topics. We therefore want to rank the results by relevance. We rank topics according to the maximum $df - idf_t$ value of their constituent terms. The motivation of this approach is assume that the most popular term from each topic represents the core of the topic. The top 10 topics are then selected for each timeslot (although this number is trivial to vary - for example, a mobile application may be designed to present fewer topics than a desktop application).

4 Data Collection

We crawled tweets for 24 hours using twitter-dataset-collector project provided by the organizers[3]. The crawler tracks four keywords (Syria, terror, Ukraine and bitcoin) and follows the users provided in the seeds.txt file.

Our final data collection is composed of 901,895 tweets and stored in Solr after filtering out the non-English tweets, extracting entities per tweet using Stanford algorithm and creating links between replies and retweets with their original tweets.

Figure 1a shows the distribution of tweets per 15 minutes timeslot for this final collection going from 18:00 25/2/2014 to 18:00 26/2/2014. The high peak detected from 20:00 to 22:00 on 25/2/2014 corresponds to sport tweets (mainly retweets and replies, see Figure 1c) related to Champions league matches that were

[3]https://github.com/socialsensor/twitter-dataset-collector

Tweets without diversity filter	*Tweets with diversity filter*
- Motorola plans to release new smartwatch this year and new version of Moto X in 'late summer' http://t.co/mLNl16fTxk by @epro - Motorola reveals it's developing a smartwatch for release in 2015 #MWC14: http://t.co/WRTYShChgH - Motorola plans to launch a smartwatch later this year http://t.co/WISyqBNgOf - In Transit From Google to Lenovo, Motorola Announces Plans For New Wearables This Year http://t.co/D7d8aXRwq4 - Motorola reveals it's developing a smartwatch to be released soon http://t.co/0OmYXDkEEr, - Motorola plans to launch a smartwatch later this year (@dcseifert / The Verge) http://t.co/QrAMc8xirX http://t.co/2zPxoOdgoS	- Motorola plans to release new smartwatch this year and new version of Moto X in 'late summer' http://t.co/mLNl16fTxk by @epro - Motorola reveals it's developing a smartwatch for release in 2015 #MWC14: http://t.co/WRTYShChgH - Motorola plans to launch a smartwatch later this year http://t.co/WISyqBNgOf - In Transit From Google to Lenovo, Motorola Announces Plans For New Wearables This Year http://t.co/D7d8aXRwq4

Table 3: Example of tweets without and with diversity filtering (using the cosine similarity of tweets) for the topic: "Motorola plans to release new smartwatch this year and new version of Moto X in late summer"

running at that time as there are some sport commentators in the newshounds list. In addition, the activity goes down at night hours as most of the newshounds are UK based, so it is more likely they were sleeping at these hours.

5 Results

The official evaluation results of our method in the Data Challenge are included in [PCA14]. Here, we summarise a few findings.

Figures 1b and 1c show the evolution of the number of tweets per timeslot containing the tracked keywords, posted by 'newshound' users, replies to 'newshounds' posts and retweets to 'newshounds' posts. Table 4 shows some representative topics associated to the tracked keywords. As can be seen in some cases, there is a strong connection between the topic timeslot and the peaks of the corresponding keyword in the chart. It confirms that our approach effectively detects topics based on the clustering of bursty terms.

In addition, our algorithm has been designed to work in real-time to keep the final users informed about the last trends in a reasonable time. According to our analysis, the BNgram approach produces a new set of topics per timeslot in 2 minutes on average, including indexing and topic detection steps. The experiments have been run in a standalone PC with 4GB RAM and an Intel CORE i3 processor.

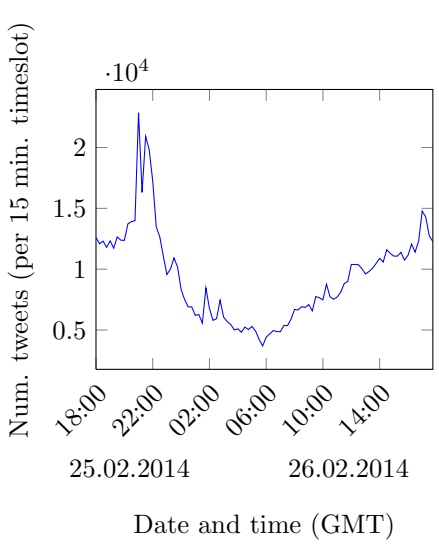
6 Discussion

After doing a first analysis of our results and previous experimentation [APM+13, MCG13, MCGM13], we know that our BNgram approach can produce a good set of trending topics based on detecting trending terms and clustering them. Here, we have applied this method to a news, challenging data set and extended the method with improved topic labelling and content diversity measures.

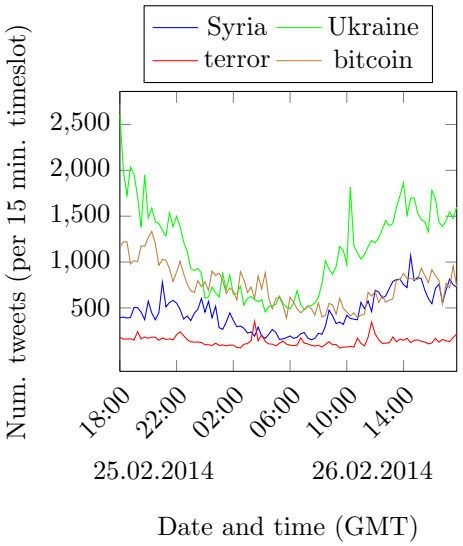
It has proven to be very challenging to determine the best label for each topic as they can easily become too specific or too general. For example, the topic label "Only one woman of color has ever won an award for Best Actress at the Academy Awards" may be too specific if the tweets within that topic mention different aspects of Academy Awards, but might be idea if that summarises what all the component tweets say. A related issue is the granularity of topics generally. To continue with this example, a higher-level story containing the same tweets and other, related tweets could be "Academy Awards results". There is no clear way to determine the "optimum" level of granularity of stories: even a brief examination of mainstream media stories shows that different stories at different levels of granularity are published all the time.

Populating topics with items based on queries could be improved with the addition of weights to the different query terms. This could be based on the $df - idf$ scores of the terms, for example. The diversity of tweets per topic is managed by the chosen threshold for cosine similarity metric. For example, a lower similarity threshold in Table 3 could reduce even more the number of diverse tweets per topic. Some further research needs to be done to compute an optimal threshold, but maybe an adaptative approach could also be considered based on different aspects of the topic (for example, number of tweets).

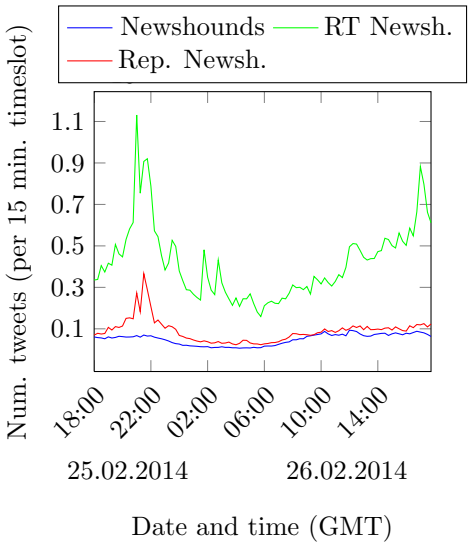
Additionally, the inclusion of replies as topic tweets tends to improve the quality of the topics (as they are not text query-dependent, so new content can be added) but it needs refinement to avoid spam tweets as discussed earlier. While replies may contain new in-



(a) Number of tweets per timeslot in the data collection



(b) Number of tweets per timeslot containing the tracking keywords



(c) Number of tweets per timeslot posted by news hounds, replies and retweets to newshounds' posts

Figure 1: Analysis of data test collection

formation, they often are merely personal responses to messages, with no newsworthy content [ZXM+]. However, the aggregation of these replies could also be useful to do sentiment analysis of users' view about the topic and detect relevant keywords in the topic which could also be helpful for the spam detection process in replies.

7 Acknowledgments

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975. Thanks to Malcolm Clark, IDEAS RGU, for all his help with the evaluation.

References

[APM+13] L.M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in Twitter. *Multimedia, IEEE Transactions on*, 15(6):1268–1282, 2013.

[AS+94] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499, 1994.

[Ben10] J. Benhardus. Streaming trend detection in Twitter. *National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, University of Colarado*, pages 1–7, 2010.

[BNG11] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11)*, 2011.

[CMG14a] David Corney, Carlos Martin, and Ayse Goker. Spot the ball: Detecting sports events on Twitter. In *European Conference on Information Retrieval ECIR2014*, pages 449–454, Amsterdam, Holland, 2014.

[CMG14b] David Corney, Carlos Martin, and Ayse Göker. Two sides to every story: Subjective event summarization of sports events using Twitter. In *ICMR2014 workshop on Social Multimedia and Storytelling*, April 2014.

[FGM05] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[MCG13] Carlos Martin, David Corney, and Ayse Goker. Finding newsworthy topics on Twitter. *IEEE Computer Society Special Technical Community on Social Networking E-Letter*, 1(3), September 2013.

[MCGM13] Carlos Martin, David Corney, Ayse Göker, and Andrew MacFarlane. Mining newsworthy topics from social media. In *BCS SGAI SMA 2013 The BCS SGAI Workshop on Social Media Analysis*, pages 35–46, 2013.

[New10] Nic Newman. #ukelection2010, mainstream media and the role of the internet. *Reuters Institute for the Study of Journalism working paper*, July 2010.

[New11] Nic Newman. Mainstream media and the distribution of news in the age of social discovery. *Reuters Institute for the Study of Journalism working paper*, September 2011.

[PCA14] Symeon Papadopoulos, David Corney, and Luca Maria Aiello. Snow 2014 data challenge: Assessing the performance of news topic detection methods in social media. In *Proceedings of the SNOW 2014 Data Challenge*, 2014.

[POL10] S. Petrovic, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Proceedings of NAACL*, volume 10, 2010.

[POL12] S. Petrovic, M. Osborne, and V. Lavrenko. Using paraphrases for improving first story detection in news and Twitter. In *Proceedings of HTL12 Human Language Technologies*, pages 338–346, 2012.

[SJ72] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[SKC11] D.A. Shamma, L. Kennedy, and E.F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 355–358. ACM, 2011.

[ZXM+] J. Zhang, Y. Xia, B. Ma, J. Yao, and Y. Hong. Thread cleaning and merging for microblog topic detection. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 589–597.

Timeslot	*Topic label*	*Keywords*	*Tweets*
		Syria	
25/2/14 20:30	Al Qaeda branch in Syria issues ultimatum to splinter group: The head of an al Qaeda-inspired militia fighting.	militia, fighting, branch, issues, ultimatum, splinter, group, inspired, head, syria, al qaeda	- Al Qaeda branch in Syria issues ultimatum to splinter group http://t.co/gQDm0p7Wur - Al Qaeda ultimatum to splinter group: The head of an al Qaeda-inspired militia fighting in Syria is giving a... http://t.co/9KFu1CG1F6
26/2/14 00:15	Jordan Bahrain Morocco Syria Qatar Oman Iraq Egypt United States 346	346	Jordan Bahrain Morocco Syria Qatar Oman Iraq Egypt United States 346 http://t.co/RjZAwwMJ95 http://t.co/RjZAwwMJ95
		terror	
26/2/14 4:00	25 marines to arrest 'worlds biggest drug lord' El Chapo Guzman 73 anti-terror-squad police to arrest 'Internet entrepreneur' Kim Dotcom	arrest, worlds, biggest, 25, marines, terror, squad, amp, police, 73, anti, drug, lord, internet, entrepren,el chapo guzman	- 25 marines to arrest 'worlds biggest drug lord' El Chapo Guzman 73 anti-terror-squad & police to arrest 'Internet entrepreneur' Kim Dotcom
		Ukraine	
26/2/14 10:15	Ukraine minister disbands Berkut riot police blamed for violence - CNN	riot, police, disbands, blamed, violence, ukraine, cnn	- RT @BBCWorld: Ukraine disbands elite Berkut anti-riot police unit, acting interior minister says http://t.co/5GqM6jjryu - RT @cnnbrk: Ukraine has disbanded a riot police force used against anti-government protesters, acting interior minister said. - RT @BBCGavinHewitt: In Ukraine the Berkut special police units blamed for most of the shootings have been disbanded.
		bitcoin	
25/2/14 20:00	Mt. Goxs Demise Marks The End of Bitcoins First Wave Of Entrepreneurs	demise, marks, end, first, wave, gox, bitcoin, entrepreneurs	- Mt. Gox's Demise Marks The End of Bitcoin's First Wave Of Entrepreneurs http://t.co/gIKKP3RLQn by @kimmaicutler - Mt. Gox's Demise Marks The End of Bitcoins First Wave Of... http://t.co/X7iUKN3Vsv #eCommerce #Finance #Startups #TC #techcrunch #tech - #SuryaRay #Surya #SuryaRay #Surya Mt. Goxs Demise Marks The End of Bitcoins... http://t.co/csX2dB26w4 @suryaray @suryaray @suryaray3

Table 4: Examples of topics about the tracking keywords