Mining Newsworthy Topics from Social Media

Carlos Martin[1], David Corney[1], Ayse Göker[1], and Andrew MacFarlane[2]

[1] IDEAS Research Institute, School of Computing & Digital Media,
Robert Gordon University, Aberdeen AB10 7QB
{c.j.martin-dancausa, d.p.a.corney, a.s.goker}@rgu.ac.uk
[2] School of Informatics, City University London, London EC1V 0HB
a.macfarlane-1@city.ac.uk

Abstract. Newsworthy stories are increasingly being shared through social networking platforms such as Twitter and Reddit, and journalists now use them to rapidly discover stories and eye-witness accounts. We present a technique that detects "bursts" of phrases on Twitter that is designed for a real-time topic-detection system. We describe a time-dependent variant of the classic *tf-idf* approach and group together bursty phrases that often appear in the same messages in order to identify emerging topics.
We demonstrate our methods by analysing tweets corresponding to events drawn from the worlds of politics and sport. We created a user-centred "ground truth" to evaluate our methods, based on mainstream media accounts of the events. This helps ensure our methods remain practical. We compare several clustering and topic ranking methods to discover the characteristics of news-related collections, and show that different strategies are needed to detect emerging topics within them. We show that our methods successfully detect a range of different topics for each event and can retrieve messages (for example, tweets) that represent each topic for the user.

Keywords: topic detection, Twitter, temporal analysis

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 [1]. 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 eye-witnesses who would be otherwise difficult to find, especially given the volume of content being shared.

Our overall goal is to produce a practical tool to help journalists and 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 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 of 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 [2, 3], where it compared favourably to several alternatives and benchmarks. Here we expand and update that work, examining the effect of different clustering and topic ranking approaches used to form coherent topics from bursty phrases.

2 Related Work

Newman [4] discusses the central use of social media by news professionals, such as hosting live blogs of ongoing events. He 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. [5] focus on the task of first-story detection (FSD), which they also call "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 neighbour and hence detect new events and track them. This work is extended in Petrovic et al. [6] 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, which is to measure the topic-level recall, i.e. to count how many newsworthy stories the system retrieved.

Benhardus [7] 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. [8] 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.

Phuvipadawat and Murata [9] analysed 154,000 tweets that contained the hashtag '#breakingnews". They determine popularity of messages by counting retweets and detecting popular terms such as nouns and verbs. This work is taken further with a simple *tf-idf* scheme that is used to identify similarity [10]; named entities are then identified using the Stanford Named Entity Recogniser in order to identify communities and similar message groups. Sayyadi et al. [11] also model the community to discover and detect events on the live Labs SocialStream platform, extracting keywords, noun phrases and named entities. Ozdikis et al. [12] also detect events using hashtags by clustering them and finding semantic similarities between hashtags, the latter being more of a lexicographic method. Ratkiewitcz et al. [13] focus specifically on the detection of a single type of topic, namely political abuse. Evidence used include the use of hashtags and mentions. Alvanaki [14] propose a system based on popular seed tags (tag pairs) which are then tracked, with any shifts detected and monitored. These articles do use natural language processing methods and most consider temporal factors, but do not use n-grams.

Becker et al. [15] 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.

3 Methods

3.1 BNgrams

Term frequency-inverse document frequency, or *tf-idf*, has been used for indexing documents since it was first introduced [16]. We are not interested in indexing documents however, but in finding novel trends, so we want to find terms that appear in one *time period* more than others. We treat temporal windows 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 (as in this work) or by following a carefully chosen list of users, and maximise *novelty* by finding bursts of suddenly high-frequency words and phrases.

We select terms with a high "temporal document frequency-inverse document frequency", or *df-idf*$_t$, by comparing the most recent x messages with the previous x messages and count how many contain the term. We regard the most recent x messages as one "slot". After standard tokenization and stop-word removal, we index all the terms from these messages. For each term, we calculate the document frequency for a set of messages using df_{ti}, defined as the number of messages in a set i that contain the term t.

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

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. To maintain some word order information, we define terms as n-grams, i.e. sequences of n words. Based on experiments reported elsewhere [3], we use 1-, 2- and 3-grams in this work. High frequency n-grams are likely to represent semantically coherent phrases. Having found bursts of potentially newsworthy n-grams, we then group together n-grams that tend to appear in the same tweets. Each of these clusters defines a topic as a list of n-grams, which we also illustrate with a representative tweet. We call this process of finding bursty n-grams"BNgrams."

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 phrases 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 phrases. We can then find the message that contains the most phrases that define a topic, and use that message as a human-readable label for the topic. We now discuss three clustering algorithms that we compare here.

Hierarchical clustering. Here, we initially assign every n-gram to its own singleton cluster, then follow a standard "group average" hierarchical clustering algorithm [17] to iteratively find and merge the closest pair of clusters. We repeat this until no two clusters share more than half their terms, at which point we assume that each cluster represents a distinct topic. We define the similarity between two terms as the fraction of messages in the same time slot that contain both of them, so it is highly likely that the term clusters whose similarities are high represent the same topic. Further details about this algorithm and its parameters can be found in our previous published work [2].

Apriori algorithm. The Apriori algorithm [18] finds all the associations between the most representative n-grams based on the number of tweets in which they co-occur. Each association is a candidate topic at the end of the process. One of the advantages of this approach is that one n-gram can belong to different associations (partial membership), avoiding one problem with hierarchical clustering. No number of associations has to be specified in advance. We also obtain maximal associations after clustering to avoid large overlaps in the final set of topic clusters.

Gaussian mixture models. GMMs assign probabilities (or strengths) of membership of each n-gram to each cluster, allowing partial membership of multiple clusters. This approach does require the number of clusters to be specified in advance, although this can be automated (e.g. by using Bayesian information criteria [19]). Here, we use the Expectation - Maximisation algorithm to optimise a Gaussian mixture model [20]. We fix the number of clusters at 20, although initial experiments showed that using more or fewer produced very similar results. Seeking more clusters in the data than there are newsworthy topics means that some clusters will contain irrelevant tweets and outliers, which can later be assigned a low rank and effectively ignored, leaving us with a few highly-ranked clusters that are typically newsworthy.

3.3 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. Here, we compare two topic ranking techniques.

Maximum *n*-gram $df - idf_t$. One method is to rank topics according to the maximum $df - idf_t$ value of their constituent n-grams. The motivation of this approach is assume that the most popular n-gram from each topic represents the core of the topic.

Weighted topic-length. As an alternative we propose weighting the topic-length (i.e. the number of terms found in the topic) by the number of tweets in the topic to produce a score for each topic. Thus the most detailed and popular topics are assigned higher rankings. We define this score thus:

$$s_t = \alpha \cdot \frac{L_t}{L_{max}} + (1 - \alpha) \cdot \frac{N_t}{N_s} \tag{2}$$

where s_t is the score of topic t, L_t is the length of the topic, L_{max} is the maximum number of terms in any current topic, N_t is the number of tweets in topic t and N_s is the number of tweets in the slot. Finally, α is a weighting term. Setting α to 1 rewards topics with more terms; setting α to 0 rewards topics with more tweets. We used $\alpha = 0.7$ in our experiments, giving slightly more weight to those stories containing more details, although the exact value is not critical.

4 Experiments

Here, we show the results of our experiments with several variations of the BNgram approach. We focus on two questions. First, what is best slot size to balance topic recall and refresh rate? A very small slot size might lead to missed stories as too few tweets would be analysed; conversely, a very large slot size means that topics would only be discovered some time after they have happened. This low 'refresh rate' would reduce the timeliness of the results. Second, what the best combination of clustering and topic ranking techniques? Earlier, we introduced three clustering methods and two topic ranking methods; we need to determine which methods are most useful.

We have previously shown that our methods perform well [2]. The BNgram approach was compared to a popular baseline system in topic detection and tracking – Latent Dirichlet Allocation (LDA) [21] – and to several other competitive topic detection techniques, getting the best overall topic recall. In addition, we have shown the benefits of using n-grams compared with single words for this sort of analysis [3]. Below, we present and discuss the results from our current experiments, starting with our approach to evaluation.

4.1 Evaluation Methods

When evaluating any IR system, it is crucial to define a realistic test problem. We used three Twitter data sets focused on popular real-world events and compare the topics that our algorithm finds with an externally-defined ground truth.

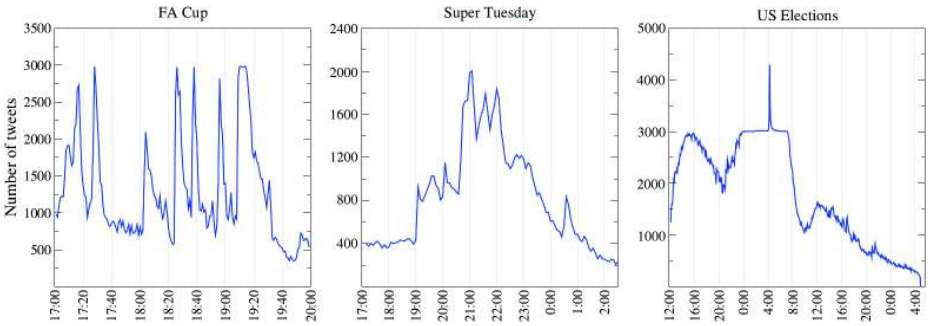

Fig. 1: Twitter activity during events (tweets per minute). For the FA Cup, the peaks correspond to start and end of the match and the goals. For the two political collections, the peaks correspond to the main result announcements.

To establish this ground truth, we relied on mainstream media (MSM) reports of the three events. This use of MSM sources helps to ensure that our ground truth topics are newsworthy (by definition) and that the evaluation is goal-focussed (i.e. will help journalists write such stories). We filtered Twitter using relevant keywords and hashtags to collect tweets around three events : the "Super Tuesday" primaries, part of the presidential nomination race of the US Republican Party; the 2012 FA Cup final, the climax to the English football season; and the 2012 US presidential election, an event of global significance. In each case, we reviewed the published MSM accounts of the events and chose a set of stories that were significant, time-specific, and represented on Twitter. For example, we ignored general reviews of the state of US politics (not time-specific), and quotes from members of the public (not significant events).

For each target topic, we identified around 5-7 keywords that defined the story to measure recall and precision, as discussed below. Some examples are shown in the first two columns of Table 4. We also defined several "forbidden" keywords. A topic was only considered as successfully recalled if all of the "mandatory" terms were retrieved and *none* of the "forbidden" terms. The aim was to avoid producing topics such as "victory Romney Paul Santorum Gingrich Alaska Georgia" that convey no information about who won or where; or "Gingrich wins", which is too limited to define the story because it doesn't name the state where the victory occurred.

Figure 1 shows the frequency of tweets collected over time, with further details in ref. [2]. We have made all the data freely available[3]. The three data sets differ in the rates of tweets, determined by the popularity of the topic and the choice of filter keywords. The mean tweets per minute (tpm) were: Super Tuesday, 832 tpm; FA Cup, 1293 tpm; and US elections, 2209 tpm. For a slot size of 1500 tweets these correspond to a "topic refresh rate" of 108s, 70s and

[3] `http://www.socialsensor.eu/results/datasets/72-twitter-tdt-dataset`

41s respectively. This means that a user interface displaying these topics could be updated every 1–2 minutes to show the current top-10 (or top-m) stories.

We ran the topic detection algorithm on each data set. This produced a ranked list of topics, each defined by a set of terms (i.e. n-grams). For our evaluation, we focus on the recall of the top m topics ($1 \leq m \leq 10$) at the time each ground-truth story emerges. For example, if a particular story was being discussed in the mainstream media from 10:00-10:15, then we consider the topic to be recalled if the system ranked it in the top m at any time during that period.

The automatically detected topics were compared to the ground truth (comprising 22 topics for Super Tuesday; 13 topics for FA Cup final; and 64 topics for US elections) using three metrics: **Topic recall**: Percentage of ground truth topics that were successfully detected by a method. A topic was considered successfully detected if the automatically produced set of words contained all mandatory keywords for it (and none of the forbidden terms, if defined). **Keyword precision**: Percentage of correctly detected keywords out of the total number of keywords for all topics detected by the algorithm in the slot. **Keyword recall**: Percentage of correctly detected keywords over the total number of ground truth keywords (excluding forbidden keywords) in the slot. One key difference between "topic recall" and "keyword recall" is that the former is a user-centred evaluation metric, as it considers the power of the system at retrieving and displaying to the user stories that are meaningful and coherent, as opposed to retrieving only some keywords that are potentially meaningless in isolation.

Note that we do not attempt to measure topic precision as this would need an estimate of the total number of newsworthy topics at any given time, in order to verify which (and how many) of the topics returned by our system were in fact newsworthy. This would require an exhaustive manual analysis of MSM sources to identify every possible topic (or some arbitrary subset), which is infeasible. One option is to compare detected events to some other source, such as Wikipedia, to verify the significance of the event [22], but Wikipedia does not necessarily correspond to particular journalists' requirements regarding newsworthiness and does not claim to be complete.

4.2 Results

Table 1 shows the effect on topic recall of varying the slot size, with the same total number of topics in the evaluation for each slot size. The mean is weighted by the number of topics in the ground truth for each set, giving greater importance to larger test sets. Overall, using very few tweets produces slightly worse results than with larger slot sizes (e.g. 1500 tweets), presumably as there is too little information in such a small collection. Slightly better results for the Super Tuesday set occur with fewer tweets; this could be due to the slower tweet rate in this set. Note that previous experiments [3] showed that including 3-grams improves recall compared to just using 1- and 2-grams, but adding 4-grams provides no extra benefit, so here we use 1-, 2- and 3-gram phrases throughout.

Slot size (tweets)	500	1000	1500	2000	2500
Super Tuesday	**0.773**	0.727	0.682	0.545	0.682
FA Cup	0.846	0.846	**0.923**	**0.923**	**0.923**
US Elections	0.750	0.781	**0.844**	0.734	0.766
Weighted mean	0.77	0.78	**0.82**	0.72	0.77

Table 1: Topic recall for different slot sizes (with hierarchical clustering).

Lastly, we compared the results of combining different clustering techniques with different topic ranking techniques (see Fig. 2). We conclude that the hierarchical clustering performs well despite the weakness discussed above (i.e. each n-gram is assigned to only one cluster), especially in FA Cup dataset. Also, the use of weighted topic-length ranking technique improves topic recall with hierarchical clustering in the political data sets.

The Apriori algorithm performs quite well in combination with the weighted topic length ranking technique (note that this ranking technique was specially created for the "partial" membership clustering techniques). We see that the Apriori algorithm in combination with the maximum n-gram $df - idf_t$ ranking technique produces slightly worse results, as this ranking technique does not produce diverse topics for the first results (from top 1 to top 10, in our case) as we mentioned earlier.

Turning to the EM Gaussian mixture model results, we see that this method works very well on the FA Cup final and US elections data sets. Despite being a "partial" membership clustering technique, the use of weighted topic length ranking technique does not make any representative difference, even its performance is worse in Super Tuesday dataset. Further work is needed to test this.

Table 2 summarises the results of the three clustering methods and the two ranking methods across all three data sets. The weighted-mean scores show that for the three clustering methods, ranking by the length of the topic is more effective than ranking by each topic's highest $df - idf_t$ score. We can see that for the FA Cup set, the Hierarchical and GMM clustering methods are the best ones in combination with the maximum n-gram $df - idf_t$ ranking technique. For Super Tuesday and US Elections data sets, "partial" membership clustering techniques (Apriori and GMM, respectively) perform the best in combination with weighted topic length ranking technique, as expected.

Finally, Table 3 shows more detailed results, including keyword precision and recall, for the best combinations of clustering and topic ranking methods of the three datasets when the top five results are considered per slot. In addition, Table 4 shows some examples of ground truth and BNgram detected topics and tweets within the corresponding detected topics for all datasets.

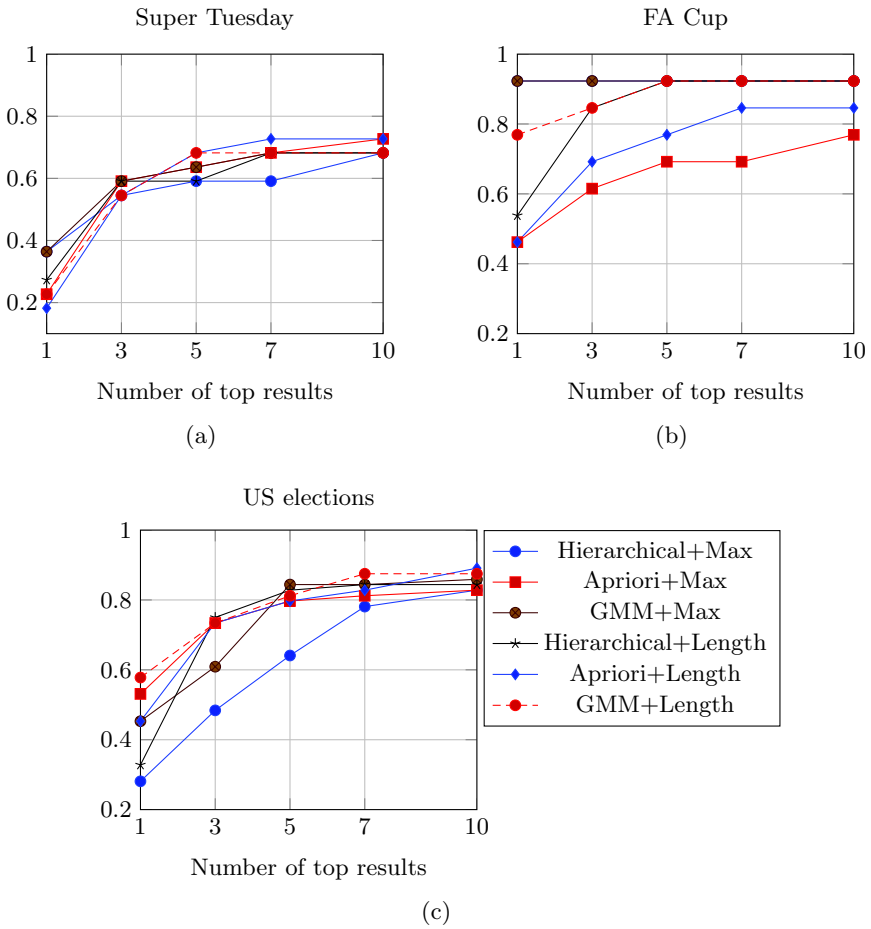

Fig. 2: Topic recall for different clustering techniques in the Super Tuesday, FA Cup and US elections sets (slot size = 1500 tweets).

5 Conclusions

If we compare the results between the three collections, one difference is particularly striking: the topic recall is far higher for football (over 90%) than for politics (around 60-80%; Table 2). This is likely to reflect the different nature of conversations about the events. Topics within a live sports event tend to be transient: fans care (or at least tweet) little about what happened five minutes ago; what matters is what is happening "now". This is especially true during key events, such as goals. In politics, conversations and comments tend to spread over hours (or even days) rather than minutes. This means that sports-related topics tend to occur over a much narrower window, with less overlapping chatter. In politics, several different topics are likely to be discussed at the same time,

Ranking	Max. n-gram $df - idf_t$			Weighted topic-length		
Clustering	Hierar.	Apriori	GMM	Hierar.	Apriori	GMM
FA Cup	**0.923**	0.677	**0.923**	0.861	0.754	0.892
Super Tuesday	0.573	0.605	0.6	0.591	**0.614**	0.586
US Elections	0.627	0.761	0.744	0.761	0.772	**0.797**
Weighted Mean	0.654	0.715	0.735	0.736	0.734	**0.763**

Table 2: Normalised area under the curve for the three datasets combining the different clustering and topic ranking techniques (1500 tweets per slot).

Method	*T-REC@5*	*K-PREC@5*	*K-REC@5*
Super Tuesday			
Apriori+Length	0.682	0.431	0.68
GMM+Length	0.682	0.327	0.753
FA Cup			
Hierar.+Max	0.923	0.337	0.582
Hierar.+Length	0.923	0.317	0.582
GMM+Max	0.923	0.267	0.582
GMM+Length	0.923	0.162	0.673
US elections			
GMM+Max	0.844	0.232	0.571

Table 3: Best results for the different datasets after evaluating top 5 topics per slot. T-REC, K-PREC, and K-REC refers to topic-recall and keyword-precision/recall respectively.

making this type of trend detection much harder. Looking back at the distribution of the tweets over time (Figure 1), we can see clear spikes in the FA Cup graph, each corresponding to a major event (kick-off, goals, half-time, full-time etc.). No such clarity is in the politics graphs, which instead is best viewed as many overlapping trends.

This difference is reflected in the way that major news stories often emerge: an initial single, focussed story emerges but is later replaced with several potentially overlapping sub-stories covering different aspects of the story. Our results suggest that a dynamic approach may be required for newsworthy topic detection, finding an initial clear burst and subsequently seeking more subtle and overlapping topics.

Recently, Twitter has been actively increasing its ties to television[4]. Broadcast television and sporting events share several common features: they occur a pre-specified times; they attract large audiences; and they are fast-paced. These features all allow and encourage audience participation in the form of sharing comments and holding discussions during the events themselves, such that the

[4] "Twitter & TV: Use the power of television to grow your impact" `https://business.twitter.com/twitter-tv`

Target topic	*Ground truth keywords*	*Extracted keywords*	*Example tweet*
Newt Gingrich says "Thank you Georgia! It is gratifying to win my home state so decisively to launch our March Momentum"	Newt Gingrich, Thank you, Georgia, March, Momentum, gratifying	launch, March, Momentum, decisively, thank, Georgia, gratifying, win, home, state, #MarchMo, #250gas, @newtgingrich	**@Bailey_Shel**: RT @newtgingrich: Thank you Georgia! It is gratifying to win my home state so decisively to launch our March Momentum. #MarchMo #250gas
Salomon Kalou has an effort at goal from outside the area which goes wide right of the goal	Salomon Kalou, run, box, mazy	Liverpool, defence, before, gets, ambushed, Kalou, box, mazy, run, @chelseafc, great, #cfcwembley, #facup, shoot	**@SharkbaitHooHa_**: RT @chelseafc: Great mazy run by Kalou into the box but he gets ambushed by the Liverpool defence before he can shoot #CFCWembley #FACup
US President Barack Obama has pledged "the best is yet to come", following a decisive re-election victory over Republican challenger Mitt Romney	Obama, best, come	America, best, come, United, States, hearts, #Obama, speech, know, victory	**@northoaklandnow**: "We know in our hearts that for the United States of America, the best is yet to come," says #Obama in victory speech.

Table 4: Examples of the mainstream media topics, the target keywords, the topics extracted by the $df\text{-}idf_t$ algorithm, and example tweets selected by our system from the collections.

focus of the discussion is constantly moving with the event itself. Potentially, this can allow targeted time-sensitive promotions and advertising based on topics currently receiving the most attention. Facebook and other social media are also competing for access to this potentially valuable "second screen" [23]. Television shows are increasingly promoting hashtags in advance, which may make collecting relevant tweets more straightforward. Even if topic detection with news requires slightly different methods compared to sport and television, both have substantial and growing demand.

Acknowledgments This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975. We wish to thank Nic Newman and Steve Schifferes of City University London for invaluable advice.

References

1. Newman, N.: Mainstream media and the distribution of news in the age of social discovery. Reuters Institute for the Study of Journalism working paper (September 2011)
2. Aiello, L., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Goker, A., Kompatsiaris, I., Jaimes, A.: Sensing trending topics in twitter. Multimedia, IEEE Transactions on **15**(6) (2013) 1268–1282
3. Martin, C., Corney, D., Goker, A.: Finding newsworthy topics on Twitter. IEEE Computer Society Special Technical Community on Social Networking E-Letter **1**(3) (September 2013)

4. Newman, N.: #ukelection2010, mainstream media and the role of the internet. Reuters Institute for the Study of Journalism working paper (July 2010)
5. Petrovic, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: Proceedings of NAACL. Volume 10. (2010)
6. Petrovic, S., Osborne, M., Lavrenko, V.: Using paraphrases for improving first story detection in news and Twitter. In: Proceedings of HTL12 Human Language Technologies. (2012) 338–346
7. Benhardus, J.: Streaming trend detection in Twitter. National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, University of Colarado (2010) 1–7
8. Shamma, D., Kennedy, L., Churchill, E.: Peaks and persistence: modeling the shape of microblog conversations. In: Proceedings of the ACM 2011 conference on Computer supported cooperative work, ACM (2011) 355–358
9. Phuvipadawat, S., Murata, T.: Breaking news detection and tracking in Twitter. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Volume 3. (2010) 120–123
10. Phuvipadawat, S., Murata, T.: Detecting a multi-level content similarity from microblogs based on community structures and named entities. Journal of Emerging Technologies in Web Intelligence **3**(1) (2011) 11–19
11. Sayyadi, H., Hurst, M., Maykov, A.: Event detection and tracking in social streams. In: Proceedings of International Conference on Weblogs and Social Media (ICWSM). (2009)
12. Ozdikis, O., Senkul, P., Oguztuzun, H.: Semantic expansion of hashtags for enhanced event detection in Twitter. In: Proceedings of VLDB 2012 Workshop on Online Social Systems. (2012)
13. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., Menczer, F.: Detecting and tracking political abuse in social media. Proc. of ICWSM (2011)
14. Alvanaki, F., Sebastian, M., Ramamritham, K., Weikum, G.: Enblogue: emergent topic detection in Web 2.0 streams. In: Proceedings of the 2011 international conference on Management of data, ACM (2011) 1271–1274
15. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on Twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM11). (2011)
16. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation **28**(1) (1972) 11–21
17. Murtagh, F.: A survey of recent advances in hierarchical clustering algorithms. The Computer Journal **26**(4) (1983) 354–359
18. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB. Volume 1215. (1994) 487–499
19. Fraley, C., Raftery, A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal **41**(8) (1998) 578–588
20. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38
21. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3** (Mar 2003) 993–1022
22. Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: First story detection using Twitter and Wikipedia. In: SIGIR 2012 Workshop on Time-aware Information Access. (2012)
23. Goel, V., Stelter, B.: Social networks in a battle for the second screen. The New York Times (October 2 2013)