SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media

Symeon Papadopoulos
CERTH-ITI
Thessaloniki, Greece
papadop@iti.gr

David Corney
Robert Gordon Univ.
Aberdeen, UK
d.p.a.corney@rgu.ac.uk

Luca Maria Aiello
Yahoo Labs
Barcelona, Spain
alucca@yahoo-inc.com

Abstract

The SNOW 2014 Data Challenge aimed at creating a public benchmark and evaluation resource for the problem of topic detection in streams of social content. In particular, given a set of tweets spanning a time interval of interest, the Challenge required the extraction of the most significant news topics in short timeslots within the selected interval. Here, we provide details with respect to the Challenge definition, the data collection and evaluation process, and the results achieved by the 11 teams that participated in it, along with a concise retrospective analysis of the main conclusions and arising issues.

1 Overview

Consider a scenario of news professionals who use social media to monitor the newsworthy stories that emerge from the crowd. The volume of information is very high and it is often difficult to extract such stories from a live social media stream. The task of the SNOW 2014 Data Challenge has been to automatically mine social streams, in particular Twitter, to provide journalists with a set of the most important topics for a number of timeslots of interest. In math terms, given a set of Twitter messages $\mathbb{M}$ spanning the interval (t_0, t_{max}) and a set of K timeslots of interest $\mathbb{S} = \{S^i | S^i = (t^i_{start}, t^i_{end})\}$, where $t^0_{start} \geq t_0$ and $t^K_{end} \leq t_{max}$, the Challenge required participants to produce K ranked lists of topics, one per timeslot: for instance, for timeslot S^i, one would produce a ranked list $T^i = \{T^i_1, T^i_2, ..., T^i_L\}$, where L is the maximum number of topics allowed per timeslot. Each topic T is associated with a headline h, a set of tags (annotations) A, a set of representative tweets $M \subset \mathbb{M}$, and optionally a set of links to images P. Table 1 summarizes the Challenge terminology.

Table 1: Challenge terminology

Symbol	Explanation
$\mathbb{M}$	Set of Twitter messages
$\mathbb{S}$	Set of timeslots of interest
$S^i = (t^i_{start}, t^i_{end}) \in \mathbb{S}$	Timeslot i
$T^i = \{T^i_1, T^i_2, ..., T^i_L\}$	Ranked list of maximum L topics for timeslot S^i
$T = (h, A, M, P)$	Topic T consists of a headline h, a set of tags A, and set of representative tweets M and images P.

The Challenge stated that the ranking of topics per timeslot should be based on the *newsworthiness* of topics. An operational definition for newsworthiness was adopted: for a given timeslot we sought topics that would turn out to be important enough to be covered in mainstream news sites.

In terms of organization, the Challenge proceeded as follows: An Open Call was published at the beginning of December 2013[1], 25 participating teams registered until the end of January, 11 successfully submitted runs at the beginning of March 2014, and 9 of them submitted papers describing their approach, making up the content of the SNOW 2014 Data Challenge proceedings.

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

[1] http://www.snow-workshop.org/2014/?page_id=37

2 Data and Ground Truth

A total of three sets of tweets were used: a development set, a rehearsal set and a test set[2]. The development set consisted of 1,106,712 tweets that we had previously collected during the 2012 US Presidential election[3]. We had previously analysed these tweets and produced a ground-truth of mainstream media stories during the period. We compared several topic-detection algorithms using this data as described in [Aie13]. The set of IDs corresponding to these tweets, with a representative sample of the associated ground-truth topics, were shared at the start of the challenge to allow participants to carry out their own internal evaluations as they developed their systems. To assist with the tweet collection, we also made available a simple tweet scraping utility[4].

For the second and third sets, we directed participants to collect tweets via the Twitter Streaming API (also making available a simple wrapper utility), filtering the stream by using provided lists of both users and keywords. By collecting tweets with the Streaming API, participants could avoid potential issues with post-hoc collection of tweets, e.g. via scraping. For the user list, we shared a previously-generated list of circa 5000 "newshounds". A newshound is a Twitter account that tends to report on and discuss major news events, and includes journalists, news outlets and agencies, commentators, politicians and other opinion-formers. In this case, the 5000 selected are UK-focussed newshounds derived initially from accounts on several public Twitter lists, and then enhanced by analysing their followers. Previous work has shown that using these newshounds to filter the Twitter stream produces a range of newsworthy tweets. Note that the Streaming API returns all messages sent by any user on the list, and also all messages that mention them. In this way, we collect messages sent to journalists, e.g. by eye-witnesses or others with information to share.

The second set of tweets was designed as a rehearsal crawl, with the aim of ensuring that the participants were able to collect and process realistic volumes of tweets. No ground truth was provided, but participants could carry out their own informal evaluations. In addition to the list of UK-focussed newshounds, we selected keywords based on events around the time of the crawl. During the rehearsal, UK news was dominated with stories of flooding in the West of England, so we used three keywords: flood, floods, and flooding.

The third set of tweets formed the test set for the final evaluation. We used the same set of UK-focussed newshounds. For the keywords, we considered which stories were likely to continue generating widespread interest and comment, making our final choice immediately before the crawl started. On the morning of the main crawl (25/02/2014), a British national and former Guantanamo Bay detainee, Moazzam Begg, had been arrested on terrorism charges related to Syria and this was likely to be discussed. The uprising in Ukraine was continuing to generate news stories and great interest. A major bitcoin exchange (Mt. Gox) had suffered a major theft, and this story was likely to generate a lot of comments online, given the technology angle of the story. We also considered protests in Venezuela, but an initial search suggested that there was relatively little UK interest in the events, possibly due to the fact that much of the first-hand reporting was in Spanish. We therefore chose four keywords: Syria, terror, Ukraine and bitcoin. The test set collection started on February 25th, 18:00 GMT and lasted for 24 hours.

This combination of keywords and newshounds was expected to produce a substantial but manageable volume of tweets, covering a wide range of stories from around the world, but of specific interest to UK-focussed journalists. In this way, we could use the UK mainstream media as a basis for deciding the "ground truth" list of target topics. It also meant that we could ignore non-English language messages and topics, avoiding the complicating issue of translation. Although it is used globally, Twitter remains dominated by English-language tweets. The final number of tweets collected was 1,041,062, representing an average of c.720 tweets per minute. Note that the exact number of tweets that each participant obtained from the Twitter Stream depends on local network connectivity and the slightly stochastic nature of the Streaming API. As we experimentally observed by running independent collection jobs, this varied by just 0.2% or 0.3% of the total number of tweets collected. We therefore shared the ID numbers of all tweets collected, allowing participants to download any tweets missing from their local collections. The tweets were sent by 556,295 accounts, contained 648,651 retweets, 135,141 replies and just 8,811 (0.85%) of them were geotagged.

We also generated the list of reference topics T^{ref} (ground truth), consisting of 59 topics that were the basis of mainstream media stories in UK news outlets during the 24-hour period of the crawl. We produced this by first collecting the headlines from both

[2]The development and test are publicly available: `http://figshare.com/articles/SNOW_2014_Data_Challenge/1003755`.

[3]The original set, also used in [Aie13], was larger, as several tweets had been removed from Twitter since the collection was first made.

[4]Source code publicly available on: `https://github.com/socialsensor/twitter-dataset-collector`

the BBC RSS news feed[5] and from NewsWhip UK[6]. From these lists, we merged duplicated stories; removed some stories of limited local or regional interest; and removed several commentary, 'op-ed' or speculative pieces. We finally checked that the remaining stories were represented in the collected tweets, and removed any that were not. This resulted in 59 distinct stories spread over 24 hours. In principle, an ideal topic-detection algorithm should be able to analyse the collection of tweets and identify all 59 stories as major news events, along with a number of other events. The aim in creating this ground-truth was not to be exhaustive (which is effectively impossible, given the scale of events taking place around the world in 24 hours, and the imprecise nature of what constitutes "news"); rather the aim is to produce a wide-ranging set of news stories covering politics, sports, international events, conflicts and so on, each of which was significant enough to generate substantial mainstream media and social media coverage.

For each story $T_i^{ref} \in T^{ref}$, we identified the approximate time that the story first appeared; the headline or label; a list of around three to five keywords or named entities defining the story; a list of two to five representative tweets from the collection; and where appropriate, the URLs of one or more related images, as shared through the collected tweets. Note that we did not expect participants to retrieve these specific tweets or URLs; they were merely indicative of the contents of the target topic. This information for each story was then used by the evaluation team to measure how effectively each participating team had discovered each story.

3 Evaluation Protocol

Participants were asked to produce up to $L = 10$ topics per 15-minute timeslot for all timeslots of the 24-hour test set interval. Thus, each participant could submit up to $24 \times 4 \times 10 = 960$ topics. The topics produced by participants were submitted in the agreed format (the same as the one used by the reference topics) to a web application. After submission, participants could browse through their topics and upload new versions of their submission until the submission deadline.

Subsequently, the evaluation was conducted by three independent evaluators, located in different countries and organizations. The web-based submission application also offered topic annotation features (cf. Figure 1) that were used to assist them in the evaluation. The evaluation was done on a set of five timeslots (starting at 18:00, 22:00, 23:15 on 25/2, and on 1:00, 1:30 on 26/2), and was blind, i.e. the evaluators did not know which participant produced any topic they evaluated. The resulting topic annotations were saved in a relational database, and aggregate statistics and results were derived with the use of SQL and some further programmatic post-processing of results in some cases.

As described in the Challenge page, four evaluation criteria were used: a) precision-recall, b) readability, c) coherence/relevance, d) diversity. The first would be quantified by means of the F-score (0-1), while the other three would be assessed on a five-level Likert scale. In the following, we provide further details with respect to the computation of the above measures. In addition, submissions were evaluated with respect to image relevance, by means of a precision score (0-1), but this was not taken into account for the final ranking since associating images with topics was optional.

3.1 Precision-recall

Precision and recall were derived with respect to two sets of reference topics: The first, T^{ref}, comprised the 59 topics manually created by the organizers as described above, while the second, denoted as T^{ext}, was created in a pooled way based on the submissions of participants during the five selected timeslots. More specifically, the evaluators assessed (using a tick box) all submitted topics during those five timeslots as being newsworthy or not (cf. paragraph 3.1.1). Topics that received at least two votes by evaluators were included in a list. After removing duplicates, a set of $|T^{ext}| = 70$ participant-pooled topics were defined. Note that a few of those topics were also included in T^{ref}.

In the case of T^{ref} we computed only recall: for each participant and for each topic of T^{ref}, the evaluators identified, with the help of a text-search facility offered by the evaluation web application, at least one matching topic in the full set of submitted topics[7]. In the end, for each participant v, we computed a recall score $R_{ref}(v) \in [0, 1]$ by dividing the number of matched topics $N_c^{ref}(v)$ with 59. Note that each evaluator performed the matching described above for a part (approximately one-third) of the 59 topics.

In the case of T^{ext}, evaluators manually matched the topics of each participant during the five selected timeslots to the topics of T^{ext}. After the matching, we could easily compute for each participant v the number of correctly matched topics $N_c^{ext}(v)$ and the number of unique correctly matched topics $N_c^{ext*}(v)$ (since a participant might detect the same topic in multiple timeslots). Then, for each participant we could com-

[5] `http://feeds.bbc.co.uk/news/rss.xml`
[6] `http://www.newswhip.com/U.K.`
[7] This was the only case where the full set of submitted topics submitted by participants was used.

pute precision and recall as follows:

$$P_{ext}(v) = \frac{N_c^{ext}(v)}{N(v)} \qquad R_{ext}(v) = \frac{N_c^{ext*}(v)}{70} \quad (1)$$

where $N(v)$ is the total number of topics submitted by v during the five selected timeslots. On the basis of precision and recall, F-score was computed as usual:

$$F_{ext}(v) = \frac{2 \cdot P_{ext}(v) \cdot R_{ext}(v)}{P_{ext}(v) + R_{ext}(v)} \quad (2)$$

3.1.1 Newsworthy assessment

Evaluators assessed each of the submitted topics (belonging to the five selected timeslots) as being newsworthy or not based on the positive and negative examples of Table 2.

Table 2: Newsworthy positive and negative examples

Type	Description
+	Major news story and/or included in T_{ref}
+	Photo-driven news story
+	Local news story
+	Announcement of future (scheduled) event
+	Goal scored in a football match
-	Opinion article
-	Analysis article
-	Speculation
-	Jokes, gossip
-	Fake news (e.g. from theonion.com)

3.2 Readability

Evaluators were instructed to assign a score between 1 and 5 (half points were also possible) according to the guidelines of Table 3. For each participant v, the readability score $Q(v)$ was computed only on the basis of the newsworthy topics, and by averaging over the three evaluators.

Table 3: Readability scoring guidelines

Score	Description
5	Understandable, readable and grammatically correct
4	Understandable but may contain minor grammatical errors
3	Includes keywords that convey the story but contains major grammatical errors
2	Hard to read and understand
1	Completely incomprehensible or nonsense

3.3 Coherence

A similar process was followed for computing coherence $C(v)$, this time using the guidelines of Table 4. The main criterion for assessing coherence is the relevance of the representative tweets with the topic headline. In addition, apart from the headline, evaluators were also instructed to consider tags: in case some of them were found to be irrelevant to the topic headline, they should decrease the coherence score (accordingly to the number of irrelevant tags). Finally, evaluators were instructed to ignore near-duplicate tweets (i.e. neither penalize nor increase the topic coherence).

Table 4: Coherence scoring guidelines

Score	Description
5	All tweets and tags are relevant
4	More relevant than non-relevant
3	About the same relevant and non-relevant
2	Less relevant than non-relevant
1	None of the tweets or tags are relevant

3.4 Diversity

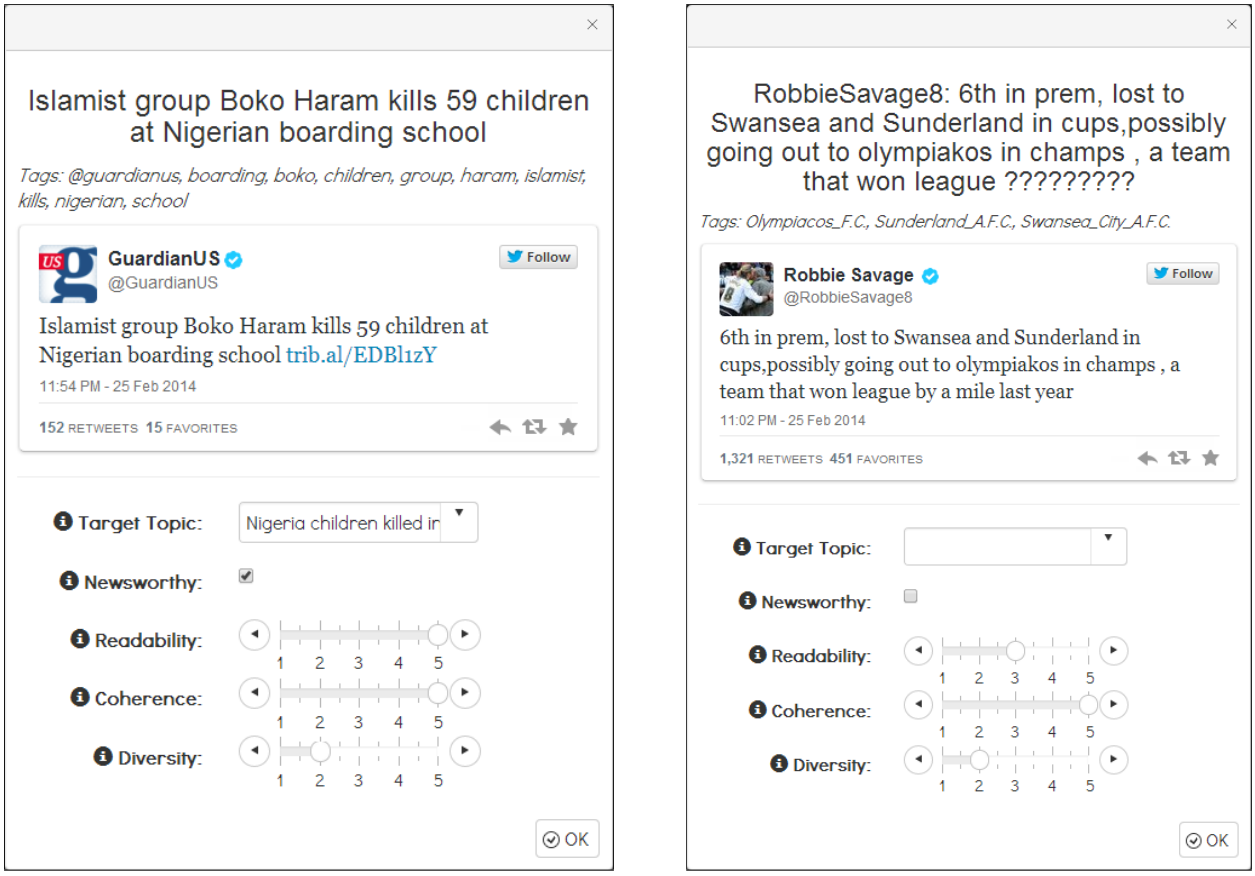
To compute diversity $D(v)$, evaluators were instructed to look into the number of *different* tweets associated with a topic: to consider a tweet as different from another, the tweet should convey some additional information. Moreover, compared to a topic that does not contain any duplication in its tweets, a topic with duplication should be slightly penalized. Depending on the degree of duplication, one may subtract 0.5 to 1 points from the score that they would otherwise assign. Table 5 provides further guidelines on assigning diversity scores.

Table 5: Diversity scoring guidelines

Score	Description
5	Several (> 3) different relevant tweets
4	A few different relevant tweets
3	At least two different relevant tweets
2	One relevant tweet
1	All tweets are irrelevant

3.5 Image relevance

To assess the relevance of an image, the evaluators needed to tick a special box in case they found the image(s) relevant to the topic under test. In cases of multiple pictures, the evaluators should make their decision based on the general impression. In the end, a single image relevance score $I(v)$ was computed for



(a) Q=5, C=5, D=2

(b) Q=3, C=5, D=2

Figure 1: Snapshot from the topic annotation interface, featuring a newsworthy and a non-newsworthy topic. Captions depict the scores assigned by one of the evaluators.

each participant by computing the percentage of relevant images in the set of newsworthy topics and averaging over the three evaluators.

3.6 Normalization and aggregation

For each of the scores used for the ranking, R_{ref}, F_{ext}, Q, C and D, we first identified the maximum attained scores R_{ref}^{max}, F_{ext}^{max}, Q^{max}, C^{max} and D^{max}, and then normalized the scores of each participant with respect to the latter. For instance, the normalized readability score would be:

$$Q^*(v) = \frac{Q(v)}{Q^{max}} \quad (3)$$

In the end, the aggregate score for each participant was derived by the following equation:

$$AS(v) = 0.25 \cdot R^*_{ref}(v) \cdot F^*_{ext}(v) + 0.25 \cdot Q^*(v) + 0.25 \cdot C^*(v) + 0.25 \cdot D^*(v) \quad (4)$$

This was the score used to derive the final ranking for the Challenge.

4 Results

Table 6 presents the raw scores achieved for each of the selected evaluation measures by the 11 participating teams. It is noteworthy that for each evaluation measure there is a different best method. For instance, the method by Insight [Ifr14] is clearly best in terms of recall (in both T^{ref} and T^{ext}) and coherence, the method by PILOTS [Nut14] best in terms of readability, while the method by SNOWBITS [Bha14] best in terms of diversity.

A second noteworthy conclusion is that almost all methods produce high-quality topic headlines ($Q > 4$) and mostly coherent topics ($C > 4$). However, the majority of methods suffer from decreased diversity ($D < 3$). This can be explained by the fact that the majority of topics produced by methods are associated with very few tweets (typically between one and three) resulting into very tight topics. Typically, while some tweets about the same topic share many of the same terms, other tweets will use distinct words. Methods that form topics based on textual-similarity of messages may therefore tend to produce these tight topics

with very low diversity. Finally, in terms of image relevance, several methods managed to achieve satisfactory performance, with more than half of the proposed images being considered as relevant to the topic headline ($I > 50\%$).

Table 9 presents the normalized scores for each criterion, the aggregate score and the final ranking for all participants. The three winning teams are Insight [Ifr14], RGU [Mar14] and math-dyn [Bur14]. One may conclude that the most distinguishing evaluation measure is topic recall and F-score with standard deviations of 0.292 and 0.29 respectively across participants, while the least discriminative measures are coherence and readability with standard deviations of 0.084 and 0.09 respectively.

Another interesting conclusion can be drawn by looking into the absolute number of unique topics that each method discovered within the five selected timeslots (Table 7). The method with the highest recall [Ifr14] managed to discover 25 of the 70 topics of T^{ext}. Given that those 70 topics are the result of topic pooling over the results of all methods, we may conclude that there is much room for improving topic recall by combining the results of multiple methods (ensemble topic discovery).

4.1 Robustness of results

To make sure that the evaluation results are robust, we looked into the following: a) inter-annotator agreement (quantified by computing the pairwise correlations of the evaluation distributions), b) alternative score aggregation methods. With respect to the first, we were pleased to note that there was significant agreement between all three evaluators across most of the evaluation aspects (readability appears to be the most subjective of all) as Table 8 testifies. With respect to the latter, we were positively surprised by the fact that several alternative normalization and aggregation schemes led to very similar rankings. More specifically, the first three methods remained the same for a number of different variations based on two schemes:

- changing the weights of the aggregation scheme of Equation 4 (instead of setting them all equal to 0.25);
- subtracting the average value for each score (instead of just dividing with the maximum value).

The stability of results over different normalization and aggregation schemes gives more confidence and credibility on the derived ranking.

Table 7: Absolute number of discovered topics

Team	N_c^{ref}	N_c^{ext}	N_c^{ext*}
UKON [Pop14]	26	13	13
IBCN [Can14]	34	12	12
ITI [Pet14]	19	22	15
math-dyn [Bur14]	37	18	14
Insight [Ifr14]	39	28	25
FUB-TORV [Ama14]	23	4	2
PILOTS [Nut14]	14	4	4
RGU [Mar14]	33	19	17
UoGMIR	10	36	15
EURECOM	14	1	1
SNOWBITS [Bha14]	8	8	7

Table 8: Inter-annotator agreement

	Eval. 1-2	**Eval. 1-3**	**Eval. 2-3**
R_{ref}	0.8949	0.9302	0.8120
P^*_{ext}	0.8956	0.8823	0.8587
Q	0.9021	0.3577	0.2786
C	0.5495	0.7307	0.6844
D	0.8734	0.8904	0.9059
I	0.9449	0.9195	0.7960

5 Outlook

In retrospect, the SNOW 2014 Data Challenge managed to bring together a number of researchers working on the problem of topic detection in noisy text streams. Conducting a fair and thorough evaluation of the competing methods proved to be a highly complicated task, calling for a variety of evaluation criteria and in-depth analysis. The results of this report along with the descriptions of the referenced methods offer a number of lessons and valuable resources to the researchers working on the field.

At this point, we should highlight a few limitations of the evaluation approach. A first one concerns the limited number of timeslots (and hence topics) assessed by the evaluators due to the limited time and resources available for evaluation. In the future, one should consider the use of crowdsourcing platforms in order to increase the breadth of the evaluation. In addition, the evaluation was limited to a specific timeslot size (15 minutes), targeting a nearly real-time scenario. Assessing the performance over larger timeslots (e.g. hour, day) could also be considered valuable for a number of applications; however, one should keep away from extrapolating the conclusions drawn from this Challenge to those settings, as the performance of different methods may be affected in different ways with the increase of timeslots (some methods might

Table 6: Overview of raw scores

Team	R_{ref}	P_{ext}	R_{ext}	F_{ext}	Q	C	D	I
UKON [Pop14]	0.44	0.481	0.186	0.268	4.29	4.40	2.12	0.542
IBCN [Can14]	0.58	0.522	0.171	0.258	4.92	4.08	2.36	0.318
ITI [Pet14]	0.32	0.440	0.214	0.288	4.49	4.68	2.31	0.581
math-dyn [Bur14]	0.63	0.462	0.200	0.279	4.59	4.91	2.11	0.520
Insight [Ifr14]	**0.66**	0.560	**0.357**	**0.436**	4.74	**4.97**	2.11	0.274
FUB-TORV [Ama14]	0.39	0.267	0.029	0.052	4.18	4.78	2.00	-
PILOTS [Nut14]	0.24	0.400	0.057	0.099	**4.93**	4.83	1.92	-
RGU [Mar14]	0.60	0.388	0.243	0.299	4.71	4.22	3.27	**0.588**
UoGMIR	0.17	**0.800**	0.214	0.338	4.80	3.95	2.36	-
EURECOM	0.24	0.125	0.014	0.027	3.38	3.75	2.50	-
SNOWBITS [Bha14]	0.14	**0.800**	0.100	0.178	4.32	4.36	**3.47**	0.186

benefit, while others might suffer).

Yet another limitation of the conducted evaluation pertains to assessing the timeliness of detected topics. When matching the submitted topics against the reference topics T^{ref}, the evaluators completely ignored the temporal information. In that way, a method that discovered a topic early on would be considered equally good with one that discovered the same topic many hours later. Obviously, this is an important performance aspect, especially in the context of breaking news detection, which should be taken into account in future evaluation efforts.

Last but not least, we should acknowledge that the type of topics sought is another important aspect for evaluating competing methods. In this Challenge, we opted for mainstream news and that was reflected in the way we constructed T^{ref}. However, by pooling results from participants (the second topic set T^{ext}), we also took into account more long-tail topics that were discovered by some of the methods. Alternative evaluation efforts may decide to give more focus on the latter, since one could argue that discovering topics that are mainstream is of limited value (except if those are discovered prior to their appearance in major news sources). Conversely, some of these long-tail topics could be popular Twitter memes or jokes that may be of limited interest to professional journalists, unless they become enormously popular.

In conclusion, the problem of topic detection is an important and attractive research topic, and the continuous increase of news-oriented social content is expected to make it even more challenging in the future. The Challenge made clear that properly assessing the performance of different methods constitutes a significant challenge on its own, and that more such efforts will be necessary in the future. For such efforts to be fruitful, the increased participation of numerous researchers working on the field is invaluable, and therefore special thanks goes to all Data Challenge participants for their hard work and patience throughout the Challenge.

Acknowledgements

This work has been supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

References

[Aie13] L. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, A. Jaimes. Sensing trending topics in Twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, Oct 2013.

[Pop14] R. Popovici, A. Weiler, M. Grossniklaus. Online Clustering for Real-Time Topic Detection in Social Media Streaming Data. *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Can14] S. Van Canneyt, M. Feys, S. Schockaert, T. Demeester, C. Develder, B. Dhoedt. Detecting Newsworthy Topics in Twitter. *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Pet14] G. Petkos, S. Papadopoulos, Y. Kompatsiaris. Two-level message clustering for topic detection in Twitter. *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Bur14] G. Burnside, D. Milioris, P. Jacquet. One Day in Twitter: Topic Detection Via Joint Complexity. *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Ifr14] G. Ifrim, B. Shi, I. Brigadir. Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. *Proceedings of SNOW 2014 Data Challenge*, 2014.

Table 9: Overview of normalized scores and aggregate results

Team	R^*_{ref}	F^*_{ext}	Q^*	C^*	D^*	I^*	AS	**Rank**
UKON [Pop14]	0.667	0.615	0.870	0.885	0.611	0.921	0.694	7
IBCN [Can14]	0.879	0.592	0.998	0.821	0.680	0.540	0.755	4
ITI [Pet14]	0.485	0.661	0.911	0.942	0.666	0.988	0.710	5
math-dyn [Bur14]	0.955	0.640	0.931	0.988	0.608	0.885	0.785	3
Insight [Ifr14]	1.000	1.000	0.961	1.000	0.608	0.466	0.892	1
FUB-TORV [Ama14]	0.591	0.119	0.848	0.962	0.576	-	0.614	10
PILOTS [Nut14]	0.364	0.227	1.000	0.972	0.553	-	0.652	9
RGU [Mar14]	0.909	0.686	0.955	0.849	0.942	1.000	0.842	2
UoGMIR	0.258	0.775	0.974	0.795	0.680	-	0.662	8
EURECOM	0.364	0.062	0.686	0.755	0.720	-	0.546	11
SNOWBITS [Bha14]	0.212	0.408	0.876	0.877	1.000	0.314	0.710	6
std. deviation	*0.292*	*0.290*	*0.090*	*0.084*	*0.146*	*0.282*	*0.100*	

[Ama14] G. Amati, S. Angelini, M. Bianchi, G. Gambosi, Gianluca Rossi. Time-based Microblog Distillation *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Nut14] G.C. Nutakki, O. Nasraoui, B. Abdollahi, M. Badami, W. Sun. Distributed LDA based Topic Modeling and Topic Agglomeration in a Latent Space. *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Mar14] C. Martin-Dancausa, A. Goker. Real-time topic detection with bursty n-grams. *Proceedings of SNOW 2014 Data Challenge*, 2014.

[Bha14] D. Bhatia, V.K. Choudhary, Y. Sharma. TwiBiNG: A Bipartite News Generator Using Twitter. *Proceedings of SNOW 2014 Data Challenge*, 2014.