Time-based Microblog Distillation

Giambattista Amati
Fondazione Ugo Bordoni
Rome, Italy
gba@fub.it

Simone Angelini
Fondazione Ugo Bordoni
Rome, Italy
sangelini@fub.it

Marco Bianchi
Fondazione Ugo Bordoni
Rome, Italy
mbianchi@fub.it

Giorgio Gambosi
Univ. of Rome Tor Vergata
Rome, Italy
giorgio.gambosi@uniroma2.it

Gianluca Rossi
Univ. of Rome Tor Vergata
Rome, Italy
gianluca.rossi@uniroma2.it

Abstract

This paper presents a simple approach for identifying relevant and reliable news from the Twitter stream, as soon as they emerge. The approach is based on a near-real time systems for sentiment analysis on Twitter, implemented by Fondazione Ugo Bordoni, and properly modified in order to detect the most representative tweets in a specified time slot. This work represents a first step towards the implementation of a prototype supporting journalists in discovering and finding news on Twitter.

1 Introduction

Microblogging is one of the most successful and widely used paradigm to communicate and interact on online social networks. According to such paradigm, users can post short messages that are publicly delivered to all their followers in real time. In particular Twitter, the most popular microblogging framework, allows to exchange messages (tweets) of most 140 chars length. This constraint is particularly suitable for posting from mobile devices, as confirmed by statistics on user access [8].

Twitter is used as a vehicle for the prompt, epidemic, diffusion of news, in terms of both announcements and comments on topics of general interest [6], though largely applied also for conversation, chatting or exchanging updates about user activities or location, that is to exchange information valuable at a personal level. With its claimed 500 million tweets per day and more than 200 million active users per month, (source: Initial public offering of shares of common stock of Twitter, Inc.) Twitter turns out as a primary source of timely information. Being able to discover relevant news as soon as they are announced, within the overall tweet stream, turns out to be an important issue both for journalists and for ordinary news readers.

This poses several non trivial problems: identifying emerging topics as collections of related tweets, recognizing news announcements from other types of information as soon as possible, determining their freshness to gather emerging news as quickly as possible, diversifying accounts of the latest news to avoid reporting the same information several times, evaluating the reliability of the news announcement also in terms of source trustfulness.

This paper reports the results of a experimentation aimed to develop a system able to effectively identify and report relevant and reliable news from the Twitter stream, as soon as they emerge. The approach is based on a near-real time system for sentiment analysis on Twitter, implemented by the Fondazione Ugo Bordoni, and properly modified in order to detect the most representative tweets in a specified time-slot.

This work represents a first step towards the implementation of a prototype supporting journalists in discovering and finding news on Twitter. To measure the effectiveness of our algorithms we have joined the SNOW 2014 Data Challenge: the task defined by orga-

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

nizers of this challenge is very suitable for our research purpose. It is worth to note, even if results of this experimentation seem to be encouraging, we consider them just a baseline for future experimentations. In fact, the effectiveness of our strategy can be improved both by a better tuning of the system parameters and by applying more advanced techniques, such as: timeline analysis to deal with freshness of tweets; sentiment analysis to detect neutrality, as expected in news announcements; more sophisticated approaches for tweet clustering and near duplicate detection.

The paper is organized as follows: in Section 2 we briefly introduce the SNOW 2014 Data Challenge task and the related benchmark. In Section 3 we provide an architectural overview of the system implemented by the Fondazione Ugo Bordoni for near-realtime sentiment analysis on Twitter. In Section 4 we describe our approach and in Section 5 we present the result of a preliminary evaluation of our baseline. Section 6 concludes the paper.

2 Task definition

The SNOW 2014 Data Challenge defines a task for real-time topic detection on Twitter. More precisely, the task consists in identifying the most relevant topics in times lots of 15 minutes in the period between 25-02-14 (18:00 GMT) and 26-02-14 (18:00 GMT).

The test data used in the SNOW 2014 Data Challenge is composed by about one million of tweets[1] from the Twitter Stream. The filtering activity has been conducted by using the Twitter Streaming API. Tweets have been selected by monitoring four keywords (i.e. Syria, terror, Ukraine, and bitcoin) and about 5000 user accounts. Since the monitoring spanned over 24 hours, the total number of analyzed time slots were 96. For each time slot and each discovered topic, a short headline should be yielded, together a set of representative tweets, possibly URLs of pictures, and finally a set of keywords. The expected output format is the following:

time-slot headline keywords tweetIds pictureUrl

With respect to the SNOW 2014 Challenge task we fulfilled the task providing the following outcomes: instead of a *headline* summarizing the discovered topic, we return the most representative tweet for that topic and we present its *tweetId* as representative tweet for the *tweetIds* field.

3 System description

The experimentation has been conducted by using a system for near-real time sentiment analysis on Twitter. This system, developed by the Fondazione Ugo Bordoni, is based on the Terrier framework [9]. Figure 1 presents an high level architectural overview of the system.

The Twitter Stream is filtered by *Twitter Connectors*, that are software components using the free Twitter Streaming API. As specified by the Twitter Streaming API Specification, each connector can define a filter composed of at most 400 keywords and 5000 user accounts. Being the usage of the API for free, the service provided by Twitter works in a best-effort fashion: as a consequence, if a filter is too much noisy (i.e. the number of tweets matching monitored keywords is too high), Twitter does not guarantees the delivery of all tweets matching conditions defined by the connector. All tweets collected by connectors are stored into a distributed installation of MongoDB [5] . Being the platform mainly oriented to implement the sentiment analysis solution described in [1], the system includes a Web application for the manual annotation of tweets and a software component (i.e. *Sentiment Analysis Dictionary Builder*) for the automatic generation of *Dictionaries* containing weighted opinion-bearing terms. Dictionaries are used by an extended version of Terrier, specifically implemented to support the indexing of tweets and to enable time-based mining activities on the indexed collection. The front-end of the system is provided by a Web application implementing several tools useful to perform time-based searches (e.g. search for relevance, search for freshness, search for opinions), to discover latent concepts related to a specified topics, providing charts, and so on. Figure 2 shows the *Buzz Chart* produced by the Web application with respect to the SNOW 2014 test collection.

This system has been used to join to the SNOW 2014 Challenge, simply submitting an "empty" query with respect to the desired time slots and setting the relevance sorting. The system automatically retrieves relevant tweets and representative weighting words for that time slots. In the following Section we detail our approach for the time-based topic distillation.

4 Experimentation

We have simulated a time-based distillation of tweets from Twitter streaming assuming that the test collection is unbiased by filtering keywords, although a very limited number of keywords were used to filter Twitter's firehose (e.g. *Syria*, *terror*, *Ukraine*, and *bitcoin*). In fact, due to this limited number of keywords, the collection can not be considered a unbiased sample of

[1] While the SNOW 2014 Data Challenge organizers collected 1.041.062 tweets, we filtered 1.040.362 tweets. Anyway the difference, in the order of 0.067%, is not statistically significant.

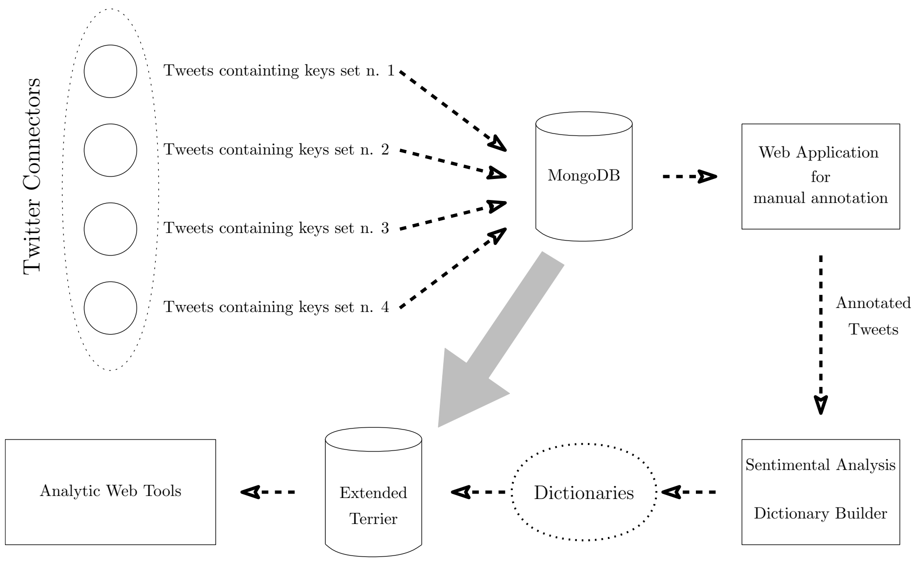

Figure 1: Architectural overview of our system for near-realtime sentiment analysis on Twitter

Twitter's firehose (about the 30% of tweets in the collection contains the above mentioned terms). As a consequence, any methodology exploiting the knowledge of these keywords could have been easily affected by overfitting. For a similar reason, and the sake of generality, we have not used the list of seed authors for filtering the news, as well as to assume ad hoc solutions for the type of task, such as the use of specific or clue keywords to detect the breaking news (e.g. the word BREAKING at the beginning of the tweets) or the id of the seed authors. However, we have deliberately removed retweets from the retrieved set because this was an explicit requirement of the task.

Since we did not have criteria or evaluation measure to assess the quality of the result set, the values of the parameters were given either by default as provided by our system or by empirical evidence. The training set thus was only used to qualitatively provide an evaluation of the distilled result set of tweets, and it was not used to tune the parameters of the filtering system.

We have submitted a run that will be used as a benchmark to evaluate our future experiments, on the basis of the evaluation measures that will be provided by the organizers [10]. For this reason we consider the submitted run just as a baseline for future experimentations.

To submit our baseline we have tackled the following issues:

a) We have assumed to process an unbiased streaming. We have gathered all tweets into time slots of 15 minutes. Thus, we have not searched tweets by using the four original topics, but we have filtered the results just by time.

b) We have used a very fast English-based filter. A set of English common terms was submitted as an unweighted query to the system and it was searched against the inverted index in order to produce a first pass retrieval. This lexicon was used to eliminate not-English tweets from the streaming. We have used a stopword list of 453 English words as a query to filter tweets written in English, and thus reducing the collection to 94.10% of the original size. The error rate of not-English tweets after retrieval was 9.03% in the sample of submitted list to the SNOW competition. We have not yet statistics on the error rate for the false negative not-retrieved set.

c) Though we have not used a query for the first pass retrieval, we have ranked the tweets of each time slot by relevance using a query expansion technique. We have applied the Bose-Einstein query expansion weights to determine the new term queries. Bose-Einstein (BO) weight is a variant of the Kullback-Leibler divergence (KL) and is preferred to KL when recall is more important than early precision, as required in our case by the absence of a topic-based first pass retrieval [3].

d) We have used a very-light and fast Near-duplicate-detection (NDD) algorithm to remove tweets from the second pass retrieved set. In particular, two tweets are considered near-duplicate if they share a bigram of two not-stopword consecutive terms. The near duplicate tweet lower in the ranking was eliminated. We finally presented the first three tweets per time slot.

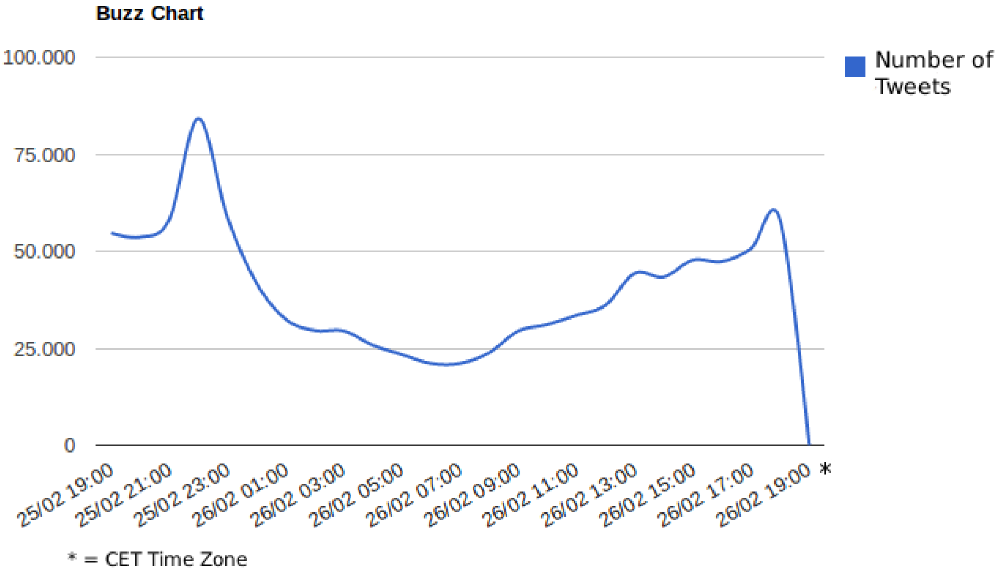

Figure 2: The Buzz chart shows the number of tweets of the SNOW 2014 test collection

5 Evaluation

Since the submitted run contains just 288 records (i.e. 3 tweets times 96 time slots), we performed a complete manual evaluation in order to evaluate the precision of our baseline. Our assessment focused on the *relevance* of tweets presented as representative for a news. Interestingly, we quickly realized it is not trivial to determine what should be classified as a news. For example:

```
115 - Sergio Aguero has the best minutes per goal
rate in @PremierLeague history scoring on average
every 115 minutes. Delivered.
```

should be considered a news? To reduce the impact of the subjectivity, each tweet has been evaluated by three human experts and classified as:

- *highly relevant* (i.e. it is a news), if all three human experts agree in considering the tweet as representative for a news;
- *not relevant* (i.e it is not a news), if all three human experts agree in considering the tweet as not representative for a news;
- *relevant* (i.e. it seems a news), otherwise.

The precision of our system (i.e. P@3) varies between 0.34, if we just consider the highly relevant class, and 0.58, if we also consider the relevant one. It is worth noting these results are strongly influenced by the choice to return exactly three tweets for each time slot. In terms of precision, this strategy can be disadvantageous when a time slot does not contain any news. Furthermore, we tried to improve the computation of the precision value and to get a sense of the performance in terms of recall: for each time slot we have extracted all tweets containing the term *BREAKING*, as substring, from the result set used to create the submitted run. In this case we assume:

- if a tweet contains the substring, it is probably a representative for a news. This hypothesis is confirmed by the resulting precision that it is equals to 0.94;
- if a news emerging from a tweet containing the substring it is not represented by a tweet in the submitted run, we missed the news;
- if a time slot does not contain any tweet containing the substring *and* all tweets in the submitted run in that time slot have been evaluated as “not relevant”, then the time slot is not considered in the computation of precision and recall values because we do not have any evidence of the existence of a news to be discovered: this hypothesis simulates the case in which the system is able to return an empty result when a time slot does not contain any news. Applying this rule 9 time slots were removed.

Considering both highly relevant and relevant classes, we obtain a precision equals to 0.64 and a recall equals to 0.80.

Even if we know we performed an incomplete assessment, we believe this is an encouraging starting point for the implementation of a vertical system for time base topic detection on Twitter. The official evalua-

Table 1: An extract of records submitted to the SNOW 2014 Challenge.

Times-lot	Headline	Tags
26-02-2014 13:30	Mt. Gox Founder Issues New Statement, Says He's Still in Japan: Mark Karpeles, founder of troubled bitcoin exc...	bitcoin,founder,new,mt,still,gox,karpeles,mark
26-02-2014 13:30	Jermain Defoe has played his last game for Tottenham after being ruled out of tomorrows match against Dnipro	last,defoe,jermain,game,dnipro,tottenham
26-02-2014 13:30	Putin puts troops in western Russia on alert amid Ukraine tension: President Vladimir Putin ordered an urgent ...	ukraine,russia,putin,troops
26-02-2014 13:45	[GMA News] Alarm in Ukraine as Putin puts Russian troops on alert: SIMFEROPOL, Ukraine/KIEV - President Vladim...	ukraine,troops,putin,russian,news,puts,alert
26-02-2014 13:45	Syria State Media Says Army Kills 175 Rebels: State media in Syria says army troops have killed 175 rebels in an ambush south of Dama...	syria,troops,state,175,media,army

Table 2: An extract of tweets containing the term BREAKING as substring.

Time-slot	Headline	Tags
26-02-2014 13:00	BREAKING: Reports say Vladimir Putin has ordered a test of combat readiness of troops in central, western Russia.”Ukraine bound!	ukraine,troops,putin,russia,combat,readiness
26-02-2014 13:30	BREAKING: State media in Syria says army troops have killed 175 rebels in an ambush south of Damascus.	syria,troops,damascus
26-02-2014 14:15	BREAKING: Mid Staffs NHS trust to be dissolved, Jeremy Hunt announces	-

tion results of our method in the Data Challenge are included in [10].

6 Conclusion and Future Work

In this paper we describe our approach in facing a challenging task: the time-based topic distillation from microblog. More precisely, we report about the strategy adopted to submit a preliminary baseline to the SNOW 2014 Data Challenge and we reported a first assessment attempt. Starting from this baseline, we will explore the following research directions:

a) The use of a topic-based clustering method, e.g. k-means driven by topic, or of a search-based result set to further split each time slot into homogeneous clusters.

b) The filtering of tweets by sentiment polarity. Sentimental analysis can be indeed useful to detect neutral tweets, since we assume that breaking news do not in general contain opinions or sentiment polarities, unless the news quotes other people's statements.

c) Freshness and tweet peak analysis improves retrieval quality [2]. The best representative for each time-based cluster can be further selected taking into account topic relevance, diversity and freshness, not just by diversity and relevance as we have done with our baseline. Zipf-law, other fat-tailed distributions [2], or exponential decaying function [7] can enhance early precision. At the moment we have not used any time-based retrieval function to order or select the tweet representatives of the selected news.

d) The NDD algorithm was very restrictive that only a few tweets were selected among the topmost relevant retrieved ones. For this reason we have decided to select only a small number of tweets per each time slot. If we had used a less aggressive Near-Duplicate Detection method, for example with Jaccard's coefficient instead of a simple bigram sharing condition, then we would have the possibility to produce a longer list of relevant and diverse news. Diversity requires thus a refinement of NDD in combination with freshness and topic relevance. Because of the too restrictive NDD condition between tweets we have not produced the list of near duplicate candidate for each selected tweet. The use of min-wise independent permutations for NDD [4] for Twitter search can be easily handled with the use of k-grams with k greater or equal to three, even without the use of sophisticated similarity functions such as Jaccard's one. In fact, due to the shortness of messages (a tweet contains 13 words on average), there is a high probability of near duplicates to share only one k-gram in a short slot of time. Obviously such tight condition would be too restrictive for larger collections and more importantly without referencing near duplicates to very short periods of time. We have thus singled out easily duplicates not only by removing the tweets containing the RT word, but also removing tweets sharing any k-gram. In order to be more selective in the initial ranking, we have further relaxed this condition to bigrams (that include entities such Mark Karpeles, western Russia etc. on Table 1), but at the moment we cannot evaluate the corresponding produced loss in recall.

7 Acknowledgments

Fondazione Ugo Bordoni carried out this work in collaboration with Almawave.

References

[1] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 89–100. Springer, 2008.

[2] G. Amati, G. Amodeo, and C. Gaibisso. Survival analysis for freshness in microblogging search. In X. wen Chen, G. Lebanon, H. Wang, and M. J. Zaki, editors, *CIKM*, pages 2483–2486. ACM, 2012.

[3] G. Amodeo, G. Amati, and G. Gambosi. On relevance, time and query expansion. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1973–1976, New York, NY, USA, 2011. Acm.

[4] A. Z. Broder and M. Mitzenmacher. Completeness and robustness properties of min-wise independent permutations. *Random Struct. Algorithms*, 18(1):18–30, 2001.

[5] K. Chodorow. *MongoDB: The Definitive Guide*. O'Reilly Media, 2013.

[6] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. *WebKDD/SNA-KDD'07*, 2007.

[7] X. Li and W. B. Croft. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 469–475, New York, NY, USA, 2003. Acm.

[8] I. Lunden. Mobile twitter: 164m+ (75%) access from handheld devices monthly, 65% of ad sales come from mobile. http://techcrunch.com/2013/10/03/mobile-twitter-161m-access-from-handheld-devices-each-month-65-of-ad-revenues-coming-from-mobile/.

[9] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier information retrieval platform. In D. E. Losada and J. M. Fernández-Luna, editors, *ECIR*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005.

[10] S. Papadopoulos, D. Corney, and L. M. 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.