Proceedings of the Ninth International AAAI Conference on Web and Social Media

Exemplar-Based Topic Detection in Twitter Streams

Ahmed Elbagoury,* **Rania Ibrahim,***
Ahmed K. Farahat, Mohamed S. Kamel, and **Fakhri Karray**
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
{ahmed.elbagoury, rania.ibrahim, afarahat, mkamel and karray}@uwaterloo.ca
∗ co-first authors

Abstract

Detecting topics in Twitter streams has been gaining an increasing amount of attention. It can be of great support for communities struck by natural disasters, and could assist companies and political parties understand users' opinions and needs. Traditional approaches for topic detection focus on representing topics using terms, are negatively affected by length limitation and the lack of context associated with tweets. In this work, we propose an Exemplar-based approach for topic detection, in which detected topics are represented using a few selected tweets. Using exemplar tweets instead of a set of key words allows for an easy interpretation of the meaning of the detected topics. Experimental evaluation on benchmark Twitter datasets shows that the proposed topic detection approach achieves the best term precision. It does this while maintaining good topic recall and running time compared to other approaches.

In 2014, active users on Twitter reached more than 645 million and produced approximately 500 million tweets daily.[1] From these numbers, users can easily miss important topics. Thus the need for mining this amount of unstructured data became paramount. Topic detection in Twitter is an important mining task that has drawn a lot of attention during the past few years. It is defined as the task of discovering the underlying key topics that occur in a set of tweets. Some of the benefits of this task include: discovering natural disasters as early as possible, helping political parties and companies understand users' opinions and improving content marketing by better understanding customer needs. Many approaches can be used to detect important topics that occur in a set of documents. However, many challenges arise when traditional approaches are applied on Twitter data. One of the challenges is the scalability of the processing methods to deal with the massive amounts of daily generated tweets.

Previous approaches that were developed in literature for topic detection like (Deerwester 1990) and (Aiello et al. 2013) focus on identifying terms that represent the topic regardless of how the terms can be properly connected so that they can be easily interpreted by an individual and regardless of whether or not noisy terms are included in the retrieved set. This motivated us to propose a fast and accurate Exemplar-based approach to detect topics in Twitter based on representing each topic by a single tweet. This Exemplar-based representation alleviates the aforementioned problems and allows for easy understanding of the retrieved topics.

The rest of the paper is organized as follows: we start first by discussing some of the related work, then presenting the proposed approach for topic detection. After that, we show the implementation details, experimental results and discussion. Finally, the last section concludes the paper.

[1] http://www.statisticbrain.com/twitter-statistics/

Related Work

A variety of techniques have been proposed for topic detection. One of them is Latent Semantic Analysis (LSA), which is a popular text analysis approach (Deerwester 1990). LSA projects a data matrix X into a lower dimensional space whose basis are latent topics. This is done by representing the matrix X as the product of three matrices ($X = U\Sigma V^T$) using Singular Value Decomposition (SVD). LSA can be performed faster using stochastic SVD (Halko, Martinsson, and Tropp 2011), hereafter referred to as stochastic LSA. Using LSA has two disadvantages: 1) The factorized matrices may have negative values which can not be easily interpreted. 2) The discovered topics are latent and do not have a clear meaning. This motivates using Non-negative Matrix Factorization (NMF) like (Lee and Seung 2001) where a matrix is factorized into the product of two matrices that don't have negative elements. We will focus on one NMF algorithm which is R1D proposed in (Biggs, Ghodsi, and Vavasis 2008). The algorithm is based on the observation that the leading singular vectors of a non-negative matrix are non-negative which yields a rank-1 approximation. R1D extends this observation to a higher rank approximation in an iterative fashion. However, NMF approaches describe each topic by a set of terms which results in topics that are not easily understood by the user and allows noisy terms to exist. Our proposed approach solves this problem by representing each topic by a real tweet (exemplar) which will not suffer from noisy terms as it is written by a human. In addition, real tweets can easily be understood, thus users can directly interpret the detected topics. Other approaches for detecting topics include using commonly used clustering methods such as K-means to cluster the set of tweets, where the tweets in the same cluster are assumed to discuss the same

topic. However, K-means clustering may result in topics being distributed over several clusters. This is solved in our approach by using exemplars (tweets) to represent the topic.

Exemplar-based Topic Detection

Notations Scalars are shown in small letters, sets are shown in script letters, vectors are denoted by small bold italic letters and matrices are denoted by capital letters. For a matrix $A \in \mathbb{R}^{n \times m}$: $A_{:i}$ denotes the i-th row of A and A_{ij} denotes the (i, j)-th entry of A. For a vector $\mathbf{x} \in \mathbb{R}^n$: $\mathbf{x_i}$ denotes the i-th element of $\mathbf{x}$

Due to the text length limitation, topic detection in short text is more challenging than in long text. So most of the existing approaches are not suitable for detecting topics in short text. The basic idea behind this work is to use Exemplar-based approach to detect topics, where each detected topic is represented using the most representative tweet. This tweet (i.e., the exemplar) is much easier to be interpreted by the user as it contains related terms and it represents a topic that is of direct importance to the user.

Problem Formulation Given a set of tweets $\mathcal{T}$ of size n, our goal is to detect the underlying topics in this set and represent each topic using only one tweet (exemplar). The selection criterion should be able to detect a tweet for each topic such that each tweet is descriptive for one topic and discriminates this topic from other topics at the same time.

Exemplar Selection Criterion The criterion used in this work is based on the following observation. A tweet which is similar to a set of tweets and dissimilar to the rest of the tweets is a good topic representative. This can be formulated by defining a similarity matrix $S_{n \times n}$ where S_{ij} is the similarity between tweet t_i and tweet t_j. The distribution of similarities between each tweet t_i and the rest of the tweets can be classified into three cases:

- Tweet t_i is similar to many tweets. Therefore, its similarity distribution will have low sample variance
- Tweet t_i is very similar to a set of tweets and less similar to the others. Therefore, its similarity distribution will have high sample variance
- Tweet t_i is not similar to most of the other tweets. So, its similarity distribution will have low sample variance

The tweets that fall in the second case are good candidates for representing topics, as each tweet is very similar to a set of tweets and therefore it can capture their underlying topic. On the other hand, each of these tweets is different from the rest of the tweets which means it can distinguish between its topic and the rest of the topics. This suggests using the variance of the similarity distribution of each tweet as a criterion for selecting topics representatives, where the sample variance of the similarities for each tweet t_i is computed as

$$\text{var}(S_{:i}) = \frac{1}{n-1} \sum_{j=1}^{n} (S_{ij} - \mu_i)^2 \,,$$

where μ_i is the mean of the similarities of tweet t_i: $\mu_i = \frac{1}{n} \sum_{j=1}^{n} S_{ij}$.

Exemplar Selection Algorithm Choosing exemplars for detected topics can be done in an iterative manner by choosing the tweet with the highest variance in each iteration as an exemplar for a topic. However, this approach is that it does not guarantee the selected tweets are talking about different topics. So after choosing each exemplar, we have to remove its effect to ensure that no more tweets about the same topic will be selected as exemplars of another topic. One way to remove the effect of an exemplar t_i is to disqualify the tweets that are ϵ close to it from being exemplars, and consider the tweet that has the highest variance of similarity and is not ϵ close to t_i as the exemplar of the next topic. Figure 1 shows an example where each node represents a tweet and its ϵ close tweets are within the dotted circle. Tweets are sorted descendingly based on the variance of their similarities with the rest of the tweets and each tweet is labeled in accordance by this order. In this example, after choosing the first tweet as the first topic exemplar, tweets 2, 3 and 4 are not chosen as exemplars of new topics as they are very close to tweet 1 and do not represent new topics. Tweet 5 which is the tweet with the highest variance of similarity and not ϵ close to tweet 1 is chosen as an exemplar of a new topic. Similarly, tweets 6 and 7 are not chosen as exemplars while tweet 8 is chosen.

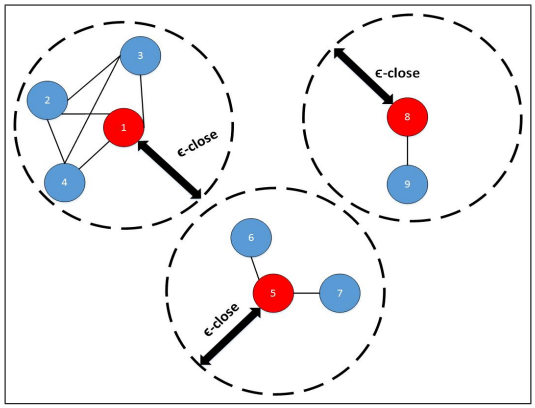

Figure 1: Exemplar-based illustration example

The set of exemplars $\mathcal{E}$ is constructed iteratively using the following objective function, at each iteration i:

$$\max_{t_i \in \mathcal{T}} \text{var}(S_{:i}), \text{ s.t. } \quad S_{ij} \leq \epsilon \quad \forall t_i, t_j \in \mathcal{E} \quad \textit{and} \quad i \neq j$$

This objective function can be solved by iterating through the tweets in descending order of the variance of their similarities and consider the first tweet that is not ϵ close to t_i as the exemplar of the next topic.

Speeding Up Calculations Computing a similarity matrix between a large number of tweets is very complex in terms of running time and memory usage. Thus to be able to handle large amounts of data, we approximate the variance of the similarities of each tweet t_i using its similarity with fewer number m of tweets, where $m < n$. So the variance of each tweet is calculated as:

$$\text{var}(S_{:i}) = \frac{1}{m-1} \sum_{j=1}^{m} (\widehat{S}_{ij} - \widehat{\mu}_i)^2,$$

Algorithm 1: Exemplar-based Topic Detection

Data: $\mathcal{T}$ set of tweets, k number of topics, m size of the random subset and ϵ similarity threshold
Result: $\mathcal{E}$ set of k tweets each representing a topic

$\widehat{\mathcal{T}} \leftarrow$ Select m random tweets from $\mathcal{T}$
$\widehat{S} \leftarrow$ similarity$(\mathcal{T}, \widehat{\mathcal{T}})$
$\widehat{\mathbf{v}} \leftarrow$ zeros(n) `// Vector of size n`
$i \leftarrow 1$
while $i \leq n$ **do**
  $\widehat{\mathbf{v}}_\mathbf{i} \leftarrow$ var$(\widehat{S}_{:i})$, $i \leftarrow i + 1$
$\widehat{\mathbf{v}} \leftarrow$ sort$(\widehat{\mathbf{v}}, \text{“}descending\text{”})$
$topic \leftarrow 1, i \leftarrow 1$
while $(topic < k)$ **do**
  $\mathcal{E}.add(\widehat{\mathbf{v}}_\mathbf{i})$, $i \leftarrow i + 1$
  while *similarity*$(\widehat{\mathbf{v}}_\mathbf{i}, \mathcal{E}(topic)) \geq \epsilon$ **do**
    $i \leftarrow i + 1$
  $topic \leftarrow topic + 1$

where $\widehat{\mu}_i = \frac{1}{m}\sum_{j=1}^{m} \widehat{S}_{ij}$ and $\widehat{S}$ is the similarity matrix between all tweets n and a random subset of size m.

It is shown empirically in the evaluation section that this approach achieves good results with better run time. The pseudo code of the approach is shown in Algorithm 1. Besides random selection, the set of m representative tweets can be selected using deterministic or hybrid techniques (Farahat et al. 2014).

Experimental Results

Two Twitter datasets were used for evaluation in this paper and were initially collected by (Aiello et al. 2013). The datasets correspond to two distinct events which include Super Tuesday (353,650 tweets) and US Elections (578,837 tweets). We have re-constructed the two datasets as only tweets' ids were provided. The ground truth topics of the datasets were constructed from news headlines reported during the events. Data is preprocessed using TMG tool[2] to remove stop words and convert it to a TF-IDF representation. The proposed topic detection approach[3] is compared against four topic detection approaches, which are: Latent Semantic Analysis (LSA), stochastic LSA (Number of iterations = 2 and oversampling = 10), Rank-1 Downdate (R1D) (Tolerance for dropping row or column = 0.1) and K-means. K-means centroids are chosen randomly and each run stops when the change in its objective function is below 0.001. Among five runs, the clusters with the minimum objective function are used. As topics in each time slot were represented by keywords, labels for each tweet were not provided. We have used the same measures and evaluation code used by (Aiello et al. 2013) to evaluate the different topic detection approaches. These measures are: topic recall (Percentage of topics successfully retrieved), term precision (Number of correct keywords in the detected topics over the total number of keywords in these detected topics) and term recall (Number of correct keywords over the total number of keywords in the ground truth topics).

Topic precision was not used by (Aiello et al. 2013) as not all the topics covered by Twitter appear in news sources. For the other topic detection approaches, the top 15 keywords of each discovered topic is used as topic keywords. While in Exemplar-based approach, each exemplar (tweet) is used as a topic and its terms are used as topic representatives. The Exemplar-based approach uses a random set to approximate the similarity matrix. Therefore, the Exemplar approach was run 10 times. The average and the 95% confidence intervals of the results were reported in Figures 2 and 3. Also, the similarity measure used was the cosine similarity and the size of the random subset of tweets m used by the Exemplar approach was set to 1000. Figures 2 and 3 show the evaluation measures of applying different topic detection approaches in Super Tuesday and US Elections datasets respectively. Exemplar-based was applied by setting the similarity threshold ϵ to 0.01 in US Elections and to 0.1 in Super Tuesday, where these values were tuned empirically. Moreover, Table 1 shows the running time for the topic detection approaches. In Super Tuesday dataset, Exemplar provided the best term precision. Moreover, Exemplar was the best in term recall and topic recall. For running time, the Exemplar and K-means approaches were the fastest. For the results of US Elections dataset, Exemplar also reached the best term precision. For topic recall, Exemplar was the best or second best in most cases, while LSA stochastic was the best in the first time slots. Term recall results were comparable for all the approaches. For the running time, the Exemplar approach and K-means were again the fastest. Achieving high term precision by the Exemplar approach is attributed to reducing the noise in the selected terms by enforcing the topics to be real tweets as shown in table 2.

Conclusion

An Exemplar-based approach is proposed for detecting topics in Twitter streams. The approach selects exemplar tweets as representatives for the detected topics based on the variance of the similarity between exemplars and other tweets. Our Exemplar-based approach achieved the best term precision as it selects real tweets as topic representatives.

Acknowledgments This publication was made possible by a grant from the Qatar National Research Fund through National Priority Research Program (NPRP) No. 06-1220-1-233. Its contents are solely the responsibility of the authors.

References

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

Biggs, M.; Ghodsi, A.; and Vavasis, S. 2008. Nonnegative matrix factorization via rank-one downdate. In *Proceedings*

[2]http://scgroup20.ceid.upatras.gr:8000/tmg/

[3]Exemplar-based code can be found at: https://sourceforge.net/projects/topicdetectionintwitterstreams/

Table 1: Running time in seconds for topic detection approaches

Dataset	# Topics (N)	R1D	LSA Stoch	LSA	K-means	Exemplar
Super Tuesday	**N = 12**	114.7866	8.3473	25.7385	**6.4527**	15.0808
	N = 92	732.1874	103.5337	387.4809	50.1507	**24.7621**
US Elections	**N = 12**	305.5000	14.4126	28.4466	**8.9827**	28.4967
	N = 92	1796.5000	209.6554	418.7506	63.8388	**45.5971**

Table 2: Sample topics detected by Exemplar-based and K-means approaches

Approach	US Elections	Super Tuesday
Exemplar	RT @AP: AP RACE CALL: Obama wins Vermont; Romney wins Kentucky. #Election2012	BREAKING NEWT: Gingrich wins Georgia Republican primary (AP)
K-means	gonna popping yeahitsope twitter live wins win obama2012 obama night election romney rt im election2012	rt romney http gingrich georgia supertuesday newt wins virginia ohio santorum mitt paul primary win

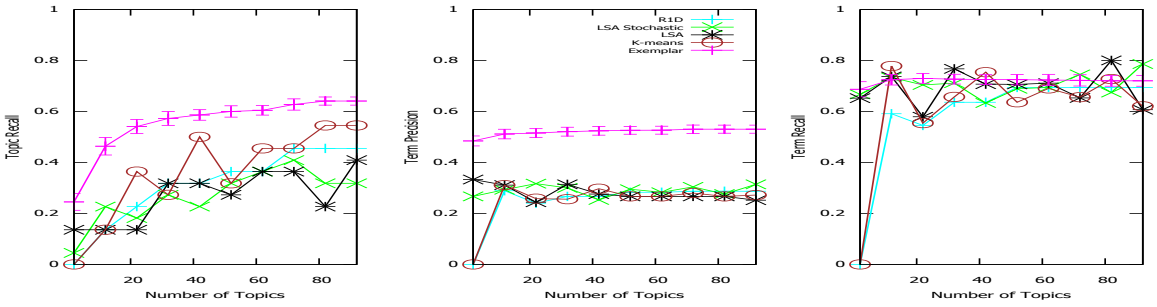

Figure 2: Results of topic detection on Super Tuesday dataset

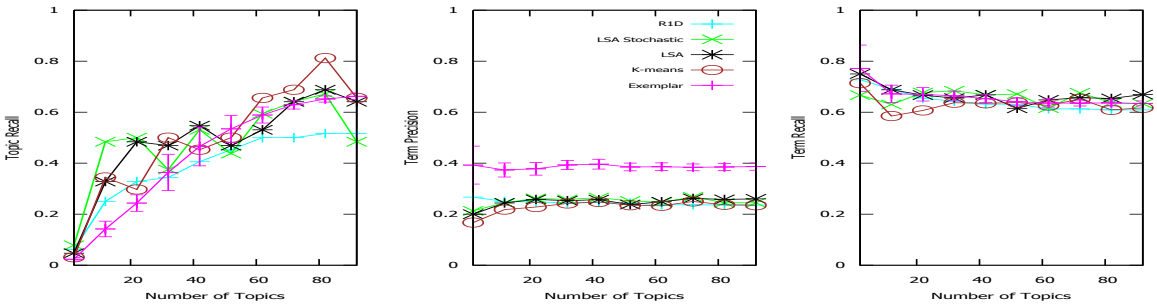

Figure 3: Results of topic detection on US Elections dataset

of the 25th international conference on Machine learning, 64–71. ACM.

Deerwester, Scott C., e. a. 1990. Indexing by latent semantic analysis. *JAsIs* 41(6):391–407.

Farahat, A.; Elgohary, A.; Ghodsi, A.; and Kamel, M. 2014. Greedy column subset selection for large-scale data sets. *Knowledge and Information Systems* 1–34.

Halko, N.; Martinsson, P.-G.; and Tropp, J. A. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53(2):217–288.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* 556–562.