Impact of Social Signals in Real Time Tweet Filtering and Summarization task

Abdelhamid Chellal, Bernard Dousset
IRIT University of Toulouse UPS, Toulouse, France
{abdelhamid.chellal,,bernard.dousset}@irit.fr

Abstract. Unlike traditional data sources, the social media stream such as Tweeter is characterized by the volume, velocity, and variety of the published information, which can vary significantly in terms of quality. Filtering and summarizing the media stream for long ongoing events is a challenging task. To be effective, a trade-off between pushing too many or too few tweets need to be achieved. While the proposed approaches are based mainly on the tweet content to discard irrelevant tweets regarding the event of interest, it is unclear how effective is the use of social signals in the tweet filtering and summarization task. We investigate the impact of social signals use to evaluate the quality of tweets. Two kinds of social signals are considered: the first ones are related to the author and the second ones are tweet specific features. Experiments were carried out on two datasets namely TREC MB Real Time Filtering 2015 and TREC MB Real Time Summarization 2016. The experiments that we conducted show the interest of integrating social signal and the use of machine learning algorithm to improve the quality of real-time tweet filtering approaches.

Keywords: Tweet filtering, Social signals, Tweet summarization.

1 Introduction

Social media stream such as Twitter is an outstanding source of information that in many cases provide real-time news before traditional media particularly those relating unexpected event such as natural disaster or terrorist attack. However, tweet stream is overwhelming, which makes it too difficult to timely follow posts describing the development of long-running events. Real-time tweet filtering is one possible solution to cope with the velocity and the volume of posted information on Twitter. This task aims to provide a means for users to keep update on topics of interest to them. In such scenario, users are looking for receiving timely and non-redundant updates as soon as a new relevant information appears in the stream.

The TREC 2015 Microblog real-time filtering (MB-TRTF)[1] and TREC 2016 real-time summarization (RTS)[2] provide experimental companies for research groups working in this area. These tracks required participants to monitor the live stream provided by Twitter over a period of ten days and to identify up to ten relevant tweets per day with respect to predefined topics (user interest). The majority of existing methods[3456] are based on query depend features to filter tweet stream. The decision to select or discard an incoming tweet depends on whether its relevance score with respect to the topic and its redundancy score fall above a predefined threshold. Although it is recognized that social signals are important for relevance, it is unclear how effective is the use of social signals in tweet filtering.

In this paper, we investigate the effect of social signals in real-time tweet filtering task by using a machine learning approach. We introduce a learn to filter approach which integrates query dependent and social signals features to build a binary classifier that produces tweet filtering predictions. Our research problem is summarized in answering the following research questions:

- What is the impact of social signals in real-time filtering tweet stream?
- How effective is learn to filter approach compared to state-of-the-art approaches?

We exploit two classes of social signals features. The first one is twitter specific features which include particular characteristics of tweets, such as retweet count and URLs. The second class is user account features which refer to the activity and the influence of the author of the post on the social network.

2 Related Work

A considerable number of methods have been proposed to filter tweet stream in real time in order to push to the user relevant tweets with respect to the topic of interest. The majority of the existing

approaches[3456] have focused on tweet content to make the decision of selection tweets for inclusion in the summary. The relevance score is evaluated using terms frequency, query term occurrence[3] in a tweet, stream statistics[45]. The TREC MB RTF-2015[1] official results reveal that runs PKUICSTRunA2[4] and UWaterlooATDK[7] are the two best performing approaches among 37 runs from 14 groups[1]. In the former, the relevance score of tweets is evaluated by using the normalized KL-divergence distance and the decision to select a tweet is based on a predefined threshold set using human intervention. In the UWaterlooATDK[7] run, the relevance score is based on the query term occurrence in the tweet. The threshold is fixed for each day according to the score of the top-10 tweets returned in the previous day. In the approach proposed in [3] authors improve the effectiveness of their approach (UWaterlooATDK[7]) by using a daily feedback strategy to estimate the relevance threshold for the next day. However, one can argue that a daily interaction for ongoing feedback judgment might be too onerous in practice. We show in this work that the result achieved by the best automatic run in TREC RTS 2016[2] in terms of precision is outperformed by the proposed approach, in which a machine learning algorithm is used.

3 Learn to filter tweet stream

To address the problem of tweet real-time filtering, we adopt a learning to filter approach which use a machine learning algorithm to build a binary classifier to produce tweet selection predictions of the incoming tweet. Learning to filter is a supervised learning task and thus has training and testing phases. The training data consists of queries which represent user interest and tweets stream. Each query is associated with a number of tweets. The relevance of the tweet with respect to the query is given. At the first, we prepare the training and test corpus using TREC MB RTF 2015 data set as described bellow. Then we extract features from the training corpus. Random Forest algorithm[8] is used to train a filter model from the training corpus. Finally, the model is evaluated by the test corpus. The binary classifier is build using Random Forest algorithm which is trained on a TREC 2015 RTF dataset and tested on the TREC 2016 RTS dataset.

3.2 Feature Description.

One of the most important tasks of machine learning algorithm is the selection of features. In the context of real-time tweet filtering, we are limited to use features that are already available in the tweet and we are not able to use Twitter REST APIs to collect further features such as the profiles of followers. We evaluated several features (around 50), we selected the top 16 using information gain algorithm implemented in Weka tool[9]. The selected features are categorized into three classes: query dependent, tweet specific and user account features.

3.2.1 Query-dependent features

Query provided in TREC MB RTF included a title of the information need and a complete description that indicates what is and is not relevant. We used five query dependent features: (1) the length of query's title $|Q^t|$, (2) the length of query's description $|Q^d|$, (3) the number of words overlap between query's title and hashtags in the tweet, (4,5) the relevance score of the incoming tweet with respect to the title Q^t and the description of the query Q^d.

Given the shortness of tweets and the streaming character of the collection, the use of statistic based approach such as Okapi BM25 appears to have limited value. Hence, the relevance score of the incoming tweet is evaluated using an adaptation of Extended Boolean Model proposed in which the word similarity based on word2vec model[10] is used to estimate the weight query terms[5]. In this approach, the query title Q^t is considered as "ANDed terms" and Q^d is considered as "ORed terms". In the Extended Boolean Model, relevance scores of the tweet $T = \{t_1,...,t_n\}$ with respect to "AND query" Q^t and "OR query" Q^d are estimated respectively as follows:

$$RSV(T,\ Q^t\ ) = 1 - \sqrt{\frac{\sum_{q_i^t \in Q^t}(1-W_T(q_i^t))^2}{|Q^t|}}\ ,\ RSV(T,Q^d) = \sqrt{\frac{\sum_{q_i^d \in Q^d}(W_T(q_i^d))^2}{|Q^d|}} \qquad (1)$$

where $W_T(q)$ is the weight of the query term q in the tweet T which is evaluated as follows:

$$W_T(q) = \max_{t_i \in T}(w2vsim(t_i, q)) \quad (3)$$

Where $w2vsim(t_i, q)$ is the similarity between tweet word t_i and query word q. Due to the fact that we cannot rely on statistics to estimate the weight of query term, we take advantage of using the vectors generated by word2vec model to evaluate the similarity between two terms. In our experiments we used using tweets crawled from 11 to 19 July 2015 (which corresponds to 9 before the TREC 2015 official evaluation periods) to generated word vectors.

3.1.2 Tweet specific features.

Tweets have many special characteristics. We exploit four tweet specific features: (i) URL& Hashtag: Whether the tweet contains an URL or a hashtag; (ii) Retweet Count per day: Ratio of times this tweet has been retweeted per the age (in day) of the tweet; (iii) number of words that a tweet contains. (iv) Whether an entity is mentioned in the tweet. For that, we use the three classes (PERSON, ORGANIZATION, LOCATION) Stanford Named Entity Recognizer[1].

3.1.3 User account features.

User account features are time-sensitive. The importance of a signal depends on the account age. An old account may have much more followers than a recent one. Therefore, in user account features, we take into account implicitly the age of the account at the time of publication tweet. The user account features are listed in table 1.

N°	Feature	Description
1	Friend	Number of friends the user has
2	Verified	Whether the user account is verified
3	Tweet/Day	Ratio of the number of tweets posted by the user per the age (in day) of the account
4	List/Day	Ratio of the number of lists a user appears in per the age (in day) of the account
5	Fol/Day	Ratio of the number of followers a user has per the age (in day) of the account
6	Fr/Day	Ratio of friends the user has per the age (in day) of the account
7	$FoD \times LD \times FrD$	combination of Fr/day, List/day and Fr/day

Table 1. User account Features

4 Experiment Data and Evaluation

4.1 Dataset

The dataset for this experiment was built from tweets captured during the evaluation period of the TREC 2015 Microblog Real Time Filtering (MB RTF) tracks, data collection was generated by each participant independently by crawling tweets using Twitter's streaming API during the evaluation period (10 days: 20 to 29 July 2015) with considering English tweets only. In this track, 51 topics were defined and the judgment pool contains 94068 tweets among them 7181 tweets were labeled by assessors as relevant with respect to one of the 51 topics. To get a balanced dataset from classes distribution point of view, we filter out all tweets that do not contain at least two query's words. Thus, we obtain a dataset that contains 6663 tweets in which the distribution of relevant and irrelevant tweet is 50.18% and 49.81% respectively.

4.2 Results and discussion

To evaluate the impact of the use of tweet specific and user account features, we experiment all features together with comparing them when only query dependent features are used alone or combined with only one class of social signals. We used 10-fold cross-validation algorithm to measure the performance of our features on TREC RTF 2015 dataset. Figure 1 reports results in terms of precision, recall, F-measure, and accuracy. As shown in figure 1, the use of social signals features improves the quality of the classifier overall metrics. We found performance improvements up to the precision, the recall and the accuracy of about 4.12%, 14.44%, and 14.46% respectively for the query dependent.

To evaluate the effectiveness of learning to filter approach, we train a binary classifier using Random Forest algorithm on a subset of TREC 2015 RTF dataset and we test this classifier by using replay

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

mechanism of scenario (A) in TREC RTS 2016 track. The training dataset consists of tweets of judgment pull that contain a least one query term. It contains 6015 relevant tweets and 39204 irrelevant tweets. In the test dataset (TREC RTS 2016), we consider only tweets that contain at least two query terms. Table 2 reports the result obtained by the binary classifier(our approach) which is compared with the high-performing automatic official results in TREC RTS 2016 track in terms of the number of relevant, redundant, irrelevant tweets and the strict precision ($P(Strict) = \frac{Rel}{Rel+Red+Not\ rel}$)..As shown in table 2, the use of binary classifier outperforms the best automatic run in TREC RTS 2016 in terms of strict precision.

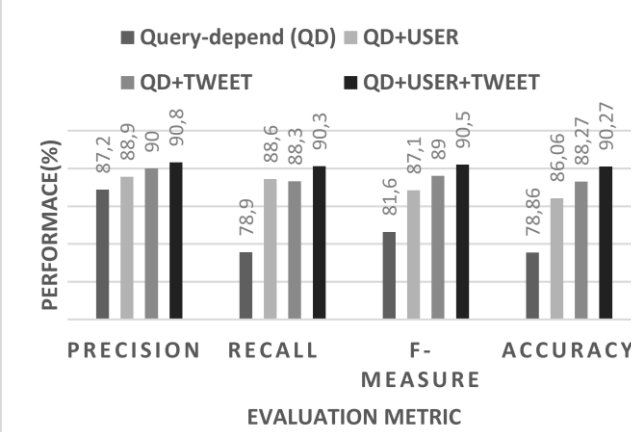

Figure 1: Performance of our features using different evaluation metrics.

Run	Rel	Red	Not Rel	P (Strict)
Our approach	91	3	84	**0.5112**
TREC 1st best automatic run	91	1	89	0.5028
TREC 2nd best automatic run	20	0	22	0.4762
TREC 3rd best automatic run	158	7	171	0.4702

Table 2 .Comparison with the official TREC 2016 RTS track results.

Conclusion

In this paper, we show that learn to filter approach that combines query dependent and social signals features can achieve good effectiveness for real-time microblog filtering. Our classifier achieves a good balance between pushing too many or too few tweets. Experiments conducted on TREC RTS 2016 reveal that the use of machine learning based filter approach produces good quality output. Our experiments highlight the importance of integrating of social signals in tweet filtering task.

References

1. J. Lin et al., Overview of the TREC 2015 Microblog Track. In Text REtrieval Conference, TREC, Gaithersburg, USA, November 17-20. (2015)
2. J.Lin et al., Overview of the TREC 2016 RealTime Summarization. In Proceedings of The Twenty-Five Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 17-20 (2016)
3. L. A. Charles et al., Simple Dynamic Emission Strategies for Microblog Filtering. In The 39th ACM SIGIR Conference (SIGIR'16). pp 745-754 (2016)
4. F. Fan, et al., PKUICST at TREC 2015 Microblog Track: Query-biased Adaptive Filtering in Real-time Microblog Stream. In Text REtrieval Conference, TREC, Gaithersburg, USA, November 17-20 (2015).
5. A.Chellal et al., Multi-criterion Real Time Tweet Summarization Based upon Adaptive Threshold. In 2016 IEEE/WIC/ACM, WI 2016, Omaha, NE, USA, October 13-16, pp 264-271 (2016).
6. A.Chellal et al., Word Similarity Based Model for Tweet Stream Prospective Notification. European Conference on Information Retrieval (ECIR 2017), Aberdeen, Scotland, UK, 08/04/17-13/04/17 (2017)
7. L.Tan, et al.,.University of Waterloo at TREC 2015 Microblog Track. In Text REtrieval Conference, Gaithersburg, USA, November 17-20 (2015)
8. L. Breiman. Random Forests. Machine Learning 45, 1 pp.5-32 (2001).
9. M. Hall et al., The weka data mining software: An update. SIGKDD Explor. Newsl., 11(1):10–18, November 2009.
10. T. Mikolov et al., Efficient Estimation of Word Representations in Vector Space. CoRR abs pp 1301-3781 (2013).