Propagated Opinion Retrieval in Twitter

Zhunchen Luo, Jintao Tang, and Ting Wang

College of Computer, National University of Defense Technology,
410073 Changsha, Hunan, China
{zhunchenluo,tangjintao,tingwang}@nudt.edu.cn

Abstract. Twitter has become an important source for people to collect opinions to make decisions. However the amount and the variety of opinions constitute the major challenge to using them effectively. Here we consider the problem of finding propagated opinions – tweets that express an opinion about some topics, but will be retweeted. Within a learning-to-rank framework, we explore a wide of spectrum features, such as retweetability, opinionatedness and textual quality of a tweet. The experimental results show the effectiveness of our features for this task. Moreover the best ranking model with all features can outperform a BM25 baseline and state-of-the-art for Twitter opinion retrieval approach. Finally, we show that our approach equals human performance on this task.

Keywords: Opinion Retrieval, Twitter, Retweet, Propagation Analysis.

1 Introduction

Twitter is the most popular micorblogging service which attracts over 500 million registered users[1] and generates over 340 million tweets daily[2]. Within Twitter, people like to share their information or opinions about personalities, politicians, products, companies, events, etc. Indeed Twitter has became an enormous repository which can not only help other people to make decisions, but also help business and government to collect valuable feedback.

However, the sheer volume of available opinions as well as the large variations present a big impediment to the effective use of the opinions in Twitter. First, the users can experience information overload due to the high volume of opinions in Twitter. Second, the importance of opinions might not be equal and the users dealing with a large number of opinions are likely to miss some important tweets. See the following tweets which are both opinions related to the topic "Obama":

(a) *"RT@KG_NYK: The fact that Obama "lost" the debate b/c he didnt call Romney's lies out well enough is pretty harrowing commentary on surf ".*
(b) *"MyNameisGurley AND I HATE OBAMA.*

[1] http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/
[2] http://blog.twitter.com/2012/03/twitter-turns-six.html

X. Lin et al. (Eds.): WISE 2013, Part II, LNCS 8181, pp. 16–28, 2013.

Users may consider tweet (a) is more important than tweet (b), since tweet (a) introduces the *First Presidential Debate* event which is related to *Obama* and gives an opinion on *Obama*'s performance. Whereas tweet (b) shows a general opinion uninteresting to most users. Moreover tweet (a) is a retweet of *KG_NYK*'s opinion by its author, which shows the agreement of the author to the original one.

Estimating the importance of a tweet is very subjective. In Twitter, however, information can deemed important by the community propagates through *retweets* [4]. This is based on human behavioral patterns for propagating microblog posts, and follows from a simple assumption: users of microblogs will propagate a post when they consider it to be important and thus worthy of being shared with other users. In this paper, we present a study of finding propagated opinions in Twitter. **Relevant tweets** should satisfy three criteria: (1) be relevant to the query; (2) contain opinions or comments about the query, irrespective of being positive or negative and (3) will be retweeted.

Previous work of predicting whether a tweet will be propagated is largely about identifying the topics of interest, and it is conceivable that unigram representation of full-length document can reasonably capture that information [4,16]. In our case, most tweets are already of interest to that user topically, which ones the user ends up retweeting may depend on several non-topical aspects of the text: whether the tweet is convincing, whether the tweet is well written, etc. Previous work has shown that such analysis can be more difficult than topic-based analysis [15], and we have the additional challenge that tweets are typically much shorter. However, the difficulty in analyzing the textural information in tweets can be alleviated by additional contextual information such as the tweets' specific information and the authors' information which potentially can improve this task.

In this paper, we use a standard machine learning approach to learn a ranking function for tweets that uses a wide spectrum of features which can recover propagated opinions in Twitter. These features include the retweetability, opinionatedness and textural quality of a tweet. The retweetability feature is the confidence score of a tweet in general being retweeted. Additionally, we proposed an approach which using social and structural information to estimating the opinionateness score of a tweet. We integrated these two features into our ranking model for this new task. Finally, we develop some features which refer to the textual quality of a tweet, including the length, the linguistic properties and the fluency of the text for a tweet. The experimental results show that the three feature sets are effective for finding propagated opinions in Twitter. Our approach integrating all feature sets performs significantly better than two baselines, one is based on the BM25 score (BM25) and the other is a state-of-the-art Twitter opinion retrieval (TOR) [13]. Moreover, a comparison of our best ranking model with human performance shows our approach does well as humans on this task.

The contributions of this paper can be summarized as follows:

1) We define a new ranking task aiming at finding opinionated tweets that will be propagated in the future.
2) We develop a set of features derived from the field of Twitter for this task and the effectiveness factors are evaluated over real-world Twitter dataset.
3) The results show the performance of our best ranking model is significantly better than the TOR baseline [13] and a BM25 baseline.
4) Furthermore, our approach for identifying the propagated opinion in Twitter can achieve human subjects' ability as well.

2 Related Work

We review related work on three main areas: message propagation and opinion mining in Twitter, review quality evaluation.

2.1 Message Propagation in Twitter

In Twitter, message deemed important by the community propagates through retweets. There is much work which is related to predicting whether a tweet in general will be retweeted. Petrovic *et al.* [16] used a machine learning approach based on the passive-aggressive algorithm to predict whether a tweet would be retweeted in the future. They found the content of the tweet, listed number, followers number and whether the author was verified were more effective features for this task. Hong *et al.* [4] proposed a method to predict the volume of retweets for a tweet. Luo [12] considered the task of finding who will retweet a message posted on Twitter. They found that followers who retweeted or mentioned the author's tweets frequently before and have common interests are more likely to be retweeters. Liu [8] investigated information propagation in Twitter from the geographical view on the global scale. They discovered that the retweet texts are more effective than common tweet texts for real-time event detection. Stieglitz [17] examined whether sentiment occurring in politically relevant tweets had an effect on their retweetability. They found a positive relationship between the quantity of words indicating affective dimensions, including positive and negative emotions associated with certain political parties or politicians, in a tweet and its retweet rate. Their work investigated whether the sentiment in a tweet could affect retweetability, but our study examines which factors affect the retweetability of opinions in Twitter.

2.2 Opinion Mining in Twitter

Twitter has attracted hundreds of millions of users who post opinions on this platform and it is also a hot research domain for academic. For example, Jansen *et al.* [5] investigated tweets as a form of electronic word-of-mouth for sharing consumer opinions concerning brands; O'Connor *et al.* [14] proposed explicitly link measurement of textural sentiment in Twitter for public opinion polls; Bollen *et al.* [1] used Twitter mood to predict the stock market, etc. However, most of these work concentrates on analyzing opinions expressed in tweets for a given topic, none on how to obtain opinions towards some persons, products or events. Luo *et al.* [13] firstly studied finding opinionated tweets for a given topic. They integrated social information and opinionatedness information into a learning to rank model. The experimental result showed that opinion retrieval performance was improved when links, mentions, author information such as the number of statues or followers and the opinionatedness of the tweet were taken into account. We take their approach as one of our baselines for comparison.

2.3 Review Quality Evaluation

Ranking reviews (opinions) according to the quality is an important problem for many online sites such as Amazon.com and Ebay.com. However, most of websites use manual votes of the *helpfulness*, such as ‘thumbs up’ and ‘thumbs down’, to assess the quality. Kim *et al.* [7] and Zhang *and* Varadarajan [19] measured the helpfulness automatically and solved it with regression model. They adopted feature sets such as lexical and syntactically oriented. The results showed that the shallow syntactic features, e.g., the counts of proper nous, modal verbs, and adjectives were correlated with the quality. Liu *et al.* [10] studied the quality of movie reviews and found, besides textural information, reviews’ expertise and the timeliness of the reviews were related to the review quality. All of these work deals with reviews in websites, Twitter, however, is a novel domain with varied short text and its rich social environment should be considered when estimating the quality.

3 Data

To investigate the factors that affect the propagation of opinions in Twitter, we use Luo *et al.* [13]’s opinion retrieval dataset[3]. It contains 50 queries and 5000 judged tweets. For each query, there are average of 16.62 opinionated tweets which are related to a given topic (query). This dataset was collected through the Twitter streaming API in November 2011. The purpose of our study is finding the opinionated tweets which will be propagated in the future. Hence, we crawled these tweets again using Twitter statuses API[4] in April 2012. Based on the principle about the relevant tweet introduced in Section 1, we take the opinionated tweets which have been retweeted within sixth months as relevant tweets and the other tweets as irrelevant tweets. We consider the state of these tweets is stable and they are not likely to be retweeted any more[5]. The task of

[3] `https://sourceforge.net/projects/ortwitter/`

[4] `https://dev.twitter.com/docs/api/1/get/statuses/show/%3Aid`

[5] When we crawled these tweets again, we found some of tweets have been deleted. We consider that if an opinionated tweet is deleted, it is not a propagated tweet any more. Therefore, we take the deleted tweets as irrelevant tweets.

this study is to show how to find these relevant tweets. The average number of relevant tweets per query is 3.4. It shows that there are only a small part of opinions which have been retweeted in Twitter and most of opinions are not be propagated. Interestingly, the percentage of opinions which have been retweeted is 20.5%, which is larger than the percentage of general tweets that have been retweeted (the value is 16.6%) in this new dataset. It shows opinions are more likely to be propagated than general tweets.

4 Overview of Our Approach

To generate a good function which ranks the tweets according our principle for finding propagated opinions in Twitter, we investigate the features concerning retweetability, opinionatedness and textural quality of a tweet. We develop a bag of features into a learning-to-rank scenario which demonstrated excellent power for ranking problem [9].

4.1 Learning to Rank Framework

Learning to rank is a data driven approach which effectively incorporates a bag of features in a model for ranking task. First, a set of queries and related tweets were used as training data. Every tweet is labeled whether it is a relevant tweet or not. A bag of features related to the relevance of a tweet is extracted to form a feature vector. Then a learning to rank algorithm is used to train a ranking model. For a new query, their related tweets, which extract the same features to form feature vectors, can be ranked by the rank function based on this model. The ranking performance of the model using a particular of feature sets in testing data can reflect the effect of these features for finding propagated opinions in Twitter.

4.2 Features for Tweets Ranking

For propagated opinion retrieval in Twitter, we consider a retweetability feature, opinionatedness feature and textural quality features for tweets ranking.

1) *Retweetability feature* refers to whether a tweet in general will be retweeted.
2) *Opinionatedness feature* refers to estimating the opinionatedness score of a tweet.
3) *Textural quality features* refer to textural information of a tweet.

In the next section, we will describe these features in details.

5 Features

5.1 Retweetability Feature

In Twitter, retweeting is an important way for information diffusion and there is a lot of work about predicting if a tweet will be retweeted [4,16]. Therefore, we develop a feature which can predict whether a tweet in general will be retweeted. We set this feature based on Petrovic *et al.* [16]. We used a machine learning approach based on the passive-aggressive algorithm to predict the retweetability score of a tweet. A set of features was developed for this prediction. It contains:

Content: the actual words in a tweet. It captures the topic of a tweet and some tweets refer to the specific topic are more likely to be retweeted. For example, people might pay more attention to the tweets related to “iran nuclear” than the tweets about “systems biology”.

Followers: the number of followers about the author of a tweet. This indicates the popularity of the user. The tweets associated with the popular authors are more likely to be retweeted.

Listed: the number of times the author of a tweet has been listed. It also indicates the popularity of the user.

Verified: whether the author of a tweet is verified. It is used by Twitter mostly to confirm the authenticity of celebrity. 91% of tweets written by verified users are retweeted, compared with 6% for tweets where the author is not verified [16].

For retweeting, the time is a critical factor. For example, people may pay more attention about the tweets related to the “American Music Awards” in November 2011 than in April 2012. Therefore, we train the prediction model on the stream of tweets crawled from the Twitter streaming API[6] throughout November 2011. We gathered a total of 30 million tweets and used them as training data. In this training data, we take the tweets which were retweeted by *retweet* button as positive samples and the other tweets as negative samples. We test the performance of our model for retweet prediction in 100,000 samples. The accuracy is 95.99%. To our **retweetability** feature, we use the margin value calculated by the passive-aggressive algorithm as the confidence of a tweet in general being retweeted.

5.2 Opinionatedness Feature

Obviously estimating the opinionatedness score of a tweet is essential for propagated opinion retrieval in Twitter. We adopt the lexicon-based approach, since it is simple and non-dependence on machine learning techniques. However, a lexicon such as *MPQA Subjectivity Lexicon*[7] which is widely used might not be effective in Twitter, since the textual content of a tweet is often very short, and lacks reliable grammatical style and quality. Therefore, we propose an approach which can automatically construct opinionated lexical from sets of tweets matching specific patterns indicative of opinionated message.

In Twitter, when people retweet another user’s tweet and give a comment before this tweet, this tweet is likely to be a subjective tweet. For example, the tweet “*I thought we were isolated and no one would want to invest here! RT @BBCNews: Honda announces 500 new jobs in Swindon bbc. in/vT12YY*” is a subjective tweet. Here, we call this tweet *Pseudo Subjective Tweet (PST).* Many

[6] http://stream.twitter.com/

[7] http://www.cs.pitt.edu/mpqa/

tweets posted by news agencies are likely to be objective tweets and these tweets usually contain links. For example, a tweet "*#NorthKorea:#KimJongil died after suffering massive heart attack on train on Saturday, official news agency reports* `bbc. in/vzPGY5`" is an objective tweet. We define a tweet satisfies two criteria: (1) it contains links and (2) the user of this tweet posted many tweets before and has many followers as *Pseudo Objective Tweet (POT).*

According to the definition introduced above, it is easy for us to design patterns and collect a large number of PSTs and POTs from Twitter. Using a PSTs set and a POTs set, we can automatically construct opinionated lexica. We use the chi-square value to estimate the opinion score of a term, which measures how dependent a term is with respect to the PSTs set and the POTs set. For the **opinionatedness feature**, we estimate the opinionatedness score of a tweet by summing all the terms with a chi-square value no less than m. The estimated formula as follows:

$$Opinion_{avg}(d) = \sum_{t \in d, \chi^2(t) \geq m} p(t|d) \cdot Opinion(t)$$

where $p(t|d) = c(t,d)/|d|$ is the relative frequency of a term t in tweet d. $c(t,d)$ is the frequency of term t in tweet d. $|d|$ is the number of terms in tweet d.

$$Opinion(t) = sgn(\frac{O_{11}}{O_{1*}} - \frac{O_{21}}{O_{2*}}) \cdot \chi^2(t)$$

where $sgn(*)$ is sign function. $\chi^2(t)$ calculates chi-square value of a term.

$$\chi^2(t) = \frac{(O_{11}O_{22} - O_{12}O_{21})^2 \cdot O}{O_{1*} \cdot O_{2*} \cdot O_{*1} \cdot O_{*2}}$$

O_{ij} in Table 1 is counted as the number of tweets having term t in the PSTs set or POTs set respectively. For example O_{12} is the number of tweets not having term t in the PSTs set.

Table 1. Table for pearson's chi-square. $O_{1*} = O_{11} + O_{12}$; $O_{2*} = O_{21} + O_{22}$; $O_{*1} = O_{11} + O_{21}$; $O_{*2} = O_{12} + O_{22}$; $O = O_{11} + O_{12} + O_{21} + O_{22}$.

	t	$\neg t$	Row total
PSTs set	O_{11}	O_{12}	O_{1*}
POTs set	O_{21}	O_{22}	O_{2*}
Column total	O_{*1}	O_{*2}	O

5.3 Textural Quality Features

Twitter is a social network that contains various content such as personal updates, babbles, conversations, etc. They are less carefully edited than other formal text (e.g., news reports) and therefore contain more misspellings and typographical errors. We develop some features which refer to the textural quality of a tweet affecting the propagation in Twitter.

Length: The total number of tokens in a tweet. Kim *et al.* [7] found the **length** feature is effective for estimating high quality reviews. Intuitively, a long tweet is apt to contain more information than a short one. We use this feature to indicate information richness for a tweet.

PosTag: Luo *et al.* [13] found the personal content is more likely to be the opinionated tweets. These tweets usually contain personal pronoun (e.g., “i”, “u” and “my”) and emotions (e.g., “:)”, “:(” and “:d”). However there is a lot of garbage which has less open-class words (i.e., nouns, verbs, adjectives and adverbs) in these tweets. E.g., the tweet *“@fayemckeever Jennifer Aniston :)”* is not a high quality opinion. Therefore we develop some features aiming to capture the linguistic properties of a tweet which include the percentage of tokens that are open-class, the percentage of tokens that are nouns, the percentage of tokens that are verbs and the percentage of tokens that are adjectives or adverbs. We use the *Twitter Part-of-Speech Tagging*[8] to tag the tweets [3].

Fluency: The fluency of a text can capture the readability of a tweet and we use language model to tackle the fluency of text. We take the probability of a tweet t under a particular language model (LM) as the fluency score $F(t)$. It is determined by:

$$F(t) = \frac{1}{m}P(w^m) = \frac{1}{m}\prod_{i}^{m} P(w_i|w_{i-N+1}, w_{i-N+2}, ..., w_{i-1})$$

where a tweet t can be expressed as a sequence of words $w^m = (w_1, w_2, ..., w_m)$. To deal with length bias, we normalize the probability by the number of tokens. We work with the N-gram based language model (N = 4) using 30 million tweets from November 2011.

6 Experiment

6.1 Human Experiments

Before estimating the performance of our approach for finding propagated opinion in Twitter, we first conduct an experiment judging whether propagated opinion can be detected by human subjects. We presented two human subjects with 100 pairs of tweets produced from our dataset, and asked them to judge which tweets were propagated opinions based on the principle introduced in Section 1. Every pair of tweets are associated to the same topic (query) and exactly one of tweets is a relevance tweet and the other is irrelevant (see the definition of relevant tweets in Section 1). The order of the two tweets in each pair was chosen randomly to avoid bias. We evaluate the performance as accuracy: the number of pairs where the human can judge which tweet is the propagated opinion correctly. In our experiment, both human subjects beat the random baseline (which is a 50% accuracy): the first subject is 75% and the other is 69%. It shows that humans are capable of judging which tweets are propagated opinions from those which are not.

[8] http://www.ark.cs.cmu.edu/TweetNLP/

6.2 Experimental Settings and Baselines

We investigate the effect of features introduced above for propagated opinion retrieval in Twitter. For learning to rank, SVM light [6] which implements the ranking algorithm is used. We use a linear kernel for training and report results for the best setting of parameters. In order to avoid overfitting the data we perform 10 fold cross-validation in our new dataset. Thus for each fold we have 45 queries with the related tweets in the training set and 5 queries with the related tweets in the testing set. We use Mean Average Precision (MAP) as the evaluation metric.

To automatically generate PSTs and POTs, we design some simple patterns: For PSTs generation, we choose the tweets uses the convention "RT @username", with text before the first occurrence of this convention. Additionally we find that the length of the preceding text should be no less than 10 characters. For POTs generation, we choose the tweets which contain a link, the author for each tweet has no less than 1,000 followers and has posted at least 10,000 tweets. In our one-month tweets dataset, 4.64% tweets are high quality PSTs and 1.35% tweets are POTs. We use 4500 PSTs and POTs[9] as opinion corpus. In our corpus-derived approach, we use the Porter English stemmer and stop words to preprocess the text of tweets. Using these tweet datasets we can calculate the value of opinionatedness score for a new tweet. To achieve the best performance of tweets ranking, we set the threshold of m is 5.02 corresponding to the significance level of 0.025 for each term in the opinion corpus. This setting is the same as [13,18].

We choose two approaches as our baselines for comparison. One is using the Okapi BM25 score of each tweet as a feature for modeling. This approach has been widely used as a baseline of Twitter retrieval [2,11,13]. We call this baseline **BM25**. The other baseline we used is based on Luo *et al.* [13]. This method integrates some social features and an opinionatedness feature for Twitter opinion retrieval. We call this baseline **TOR**. The detail of the features in **TOR** baseline are shown in Table 2.

Table 2. TOR Baseline Features

TOR Features	**Description**
BM25	The Okapi BM25 score
Mention	A binary feature whether a tweet contains "@username"
URL	A binary feature whether a tweet contains a link
Statuses	The number of tweets (statuses) the author has ever written
Followers	The number of followers the author has
Opinionatedness	The opinionatedness score of a tweet

[9] We test that 4500 PSTs and POTs as corpus for estimating opinionatedness feature can achieve high performance for propagated opinion ranking in Twitter and there is no significant improvement adding more PSTs and POTs.

6.3 Result

We investigate whether the features introduced in Section 5 are effective for propagated opinion retrieval in Twitter. We integrate each feature with the two baselines features into our tweets ranking systems respectively. Table 3 and Table 4 show the performance of each ranking model.

We can see that using **Retweetability**, **PosTag** and **Fluency** features significantly improve the results when integrated with the **TOR**. It suggests the retweetability, the linguistic properties and the readability of a tweet can indeed help finding propagated opinions in Twitter. We can also see that the performance **TOR** is significantly better than **BM25** ($p<0.01$). It is not surprising that the opinionatedness information and some social information of tweets are essential for this task. Although the performance results of the **BM25** ranking model integrated with **Retweetability**, **PosTag** and **Fluency** respectively are higher than **BM25**, they are not significant. The reason may be that just using these features alone are not enough for improving tweets ranking. Interestingly, we find the **Length** feature can help finding propagated opinion integrated with **TOR**, but the result is decreased combined with the **BM25**. It shows the length information is not very effective for finding propagated opinions in Twitter as other review websites [7]. This is because each tweet has to follow the 140-characters limitation, therefore the diversity of length between propagated opinions and the other tweets is not obvious. We integrate the **Textual Quality** features (combine **Length**, **PosTag** and **Fluency** together) into the two baselines and find the performance is improved more. All these show that our **Retweetability**, **Opinionatedness** and **Textual Quality** are all effective for finding propagated opinions in Twitter.

Table 3. BM25 is a baseline. A significantly improvement with $^{\triangle}$ and $^{\blacktriangle}$ (for $p < 0.05$ and $p < 0.01$ respectively). BM25+All combines BM25, Retweetability, Opinionatedness and Textural Quality features together.

	MAP
BM25	0.0997
BM25+Retweetability	0.1077
BM25+Opinionatedness	0.1146
BM25+Length	0.0881
BM25+PosTag	0.1157
BM25+Fluency	0.1046
BM25+Textural Quality	0.1277
BM25+All	0.1317

At last we add all the features based on **TOR** baseline into a ranking model (**TOR+Retweetability+Textural Quality**). Table 4 shows its best result achieved the MAP value 0.1992. The best result improves MAP by 30.97% over the **TOR** method and 99.80% over the **BM25** method. All these show our **Best** ranking model can not only find the opinionated tweets to a given topic, but these tweets are also more likely to be propagated in the future. For example, the query *American Music Awards* yields three tweets in our data:

(a) *Watch Olnine Free— The 39th Annual American Music Awards (TV 2011): The 39th Annual American Music Awards (TV 20...* `http://t.co/SxrjVVmx`.
(b) *We're so excited for the American Music Awards this weekend.*
(c) *That awkward moment when the American Music Awards is really the American Minaj Awards.*

In our experiment, the **BM25** method ranks tweet (a) higher than tweet (b) and tweet (c), but this tweet is an objective message without opinions. **TOR** ranks tweet (b) higher than the other tweets, since it contains the author's opinion about the *American Music Awards*, however it was not propagated. Our **Best** ranking model ranks tweet (c) higher and this funny opinion had been propagated 143 times within six months.

Table 4. TOR is a baseline. A significantly improvement with △ and ▲ (for p < 0.05 and p < 0.01 respectively).

	MAP
TOR	0.1521
TOR+Retweetability	0.1806▲
TOR+Length	0.1580
TOR+PosTag	0.1917▲
TOR+Fluency	0.1875△
TOR+Textural Quality	0.1930▲
TOR+Retweetability+Textural Quality (Best)	0.1992▲

6.4 Opinion Propagation Prediction vs General Message Propagation Prediction

There are much work which predicts whether a tweet in general will be retweeted [4,16]. We are interested in the relationship of propagation predictions between opinions and general message in Twitter. We investigate whether only using the **Retweetability** feature is enough to find the propagated opinionated tweets. Table 5 gives the result that the performance of **Retweetability** ranking model is worse than **Best** ranking model significantly. It shows the task of predicting whether an opinion will be propagated is different to the related task of predicting whether a tweet in general will be propagated. Therefore, to the task in this study, we should consider more information such as the opinionatedness and textual quality of tweets.

6.5 Comparison with Humans

Finally, using the **Best** ranking model for finding propagated opinions in Twitter, we turn back to see human experiment in Section 6.1. We use our ranking

Table 5. Retweetability is a baseline. A significant improvement with $^{\triangle}$ and $^{\blacktriangle}$ (for p < 0.05 and p < 0.01 respectively).

	MAP
Retweetability	0.0936
TOR+Retweetability+Textural Quality (Best)	$0.1992^{\blacktriangle}$

model to judge which tweets presented are more likely to be propagated opinions. This model achieved an accuracy of 71%, which is slightly lower than human subjects (average 72%), but not significantly different from either subject at p=0.05. This result shows that for the task of finding propagated opinion in Twitter our approach is able to do as well as humans.

7 Conclusion

In this paper we study the task aiming at finding propagated opinions in Twitter. A set of features, including the retweetability, opinionatedness and textural quality of a tweet, are developed and integrated into learning to rank model for solving this task. The experimental results show these features are effective for finding propagated opinions in Twitter. Moreover, our best ranking model integrating all features is significantly better than the start-of-the-art TOR baseline and a BM25 baseline. Finally, we are encouraged by the performance of our ranking model, which can achieve the human subjects' ability as well, in identifying the propagated opinions in Twitter.

Acknowledgements. We would like to thank Zheng Wang for tagging the data. This research is supported by the National Natural Science Foundation of China (Grant No. 61170156 and 61202337).

References

1. Bollen, J., Mao, H., Zeng, X.-J.: Twitter mood predicts the stock market. J. Comput. Science 2(1), 1–8 (2011)
2. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.Y.: An empirical study on learning to rank of tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, pp. 295–303. Association for Computational Linguistics, Stroudsburg (2010)
3. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, HLT 2011, vol. 2, pp. 42–47. Association for Computational Linguistics, Stroudsburg (2011)
4. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web, WWW 2011, pp. 57–58. ACM, New York (2011)

5. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. 60(11), 2169–2188 (2009)
6. Joachims, T.: Making large scale svm learning practical (1999)
7. Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically assessing review helpfulness. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 423–430. Association for Computational Linguistics (2006)
8. Liu, P., Tang, J., Wang, T.: Information current in twitter: which brings hot events to the world. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 111–112. International World Wide Web Conferences Steering Committee (2013)
9. Liu, T.Y.: Learning to rank for information retrieval. Found. Trends Inf. Retr. 3(3), 225–331 (2009)
10. Liu, Y., Huang, X., An, A., Yu, X.: Modeling and predicting the helpfulness of online reviews. In: ICDM, pp. 443–452 (2008)
11. Luo, Z., Osborne, M., Petrovic, S., Wang, T.: Improving twitter retrieval by exploiting structural information. In: AAAI 2012: Proceedings of the Twenty-Sixth AAAI (2012)
12. Luo, Z., Osborne, M., Tang, J., Wang, T.: Who will retweet me? finding retweeters in twitter. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2013)
13. Luo, Z., Osborne, M., Wang, T.: Opinion retrieval in twitter. In: ICWSM (2012)
14. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: ICWSM (2010)
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. 2(1-2), 1–135 (2008)
16. Petrovic, S., Osborne, M., Lavrenko, V.: Rt to win! predicting message propagation in twitter. In: ICWSM (2011)
17. Stieglitz, S., Dang-Xuan, L.: Political communication and influence through microblogging-an empirical analysis of sentiment in twitter messages and retweet behavior. In: HICSS, pp. 3500–3509 (2012)
18. Zhang, W., Yu, C., Meng, W.: Opinion retrieval from blogs. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM 2007, pp. 831–840. ACM, New York (2007)
19. Zhang, Z., Varadarajan, B.: Utility scoring of product reviews. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 51–57. ACM (2006)