World Wide Web (2015) 18:545–566
DOI 10.1007/s11280-013-0268-7

An effective approach to tweets opinion retrieval

Zhunchen Luo · Miles Osborne · Ting Wang

Received: 3 January 2013 / Revised: 18 September 2013 /
Accepted: 31 October 2013 / Published online: 6 December 2013

Abstract Opinion retrieval deals with finding relevant documents that express either a negative or positive opinion about some topic. Social Networks such as Twitter, where people routinely post opinions about almost any topic, are rich environments for opinions. However, spam and wildly varying documents makes opinion retrieval within Twitter challenging. Here we demonstrate how we can exploit social and structural textual information of Tweets and improve Twitter-based opinion retrieval. In particular, within a learning-to-rank technique, we explore the question of whether aspects of an author (such as the number of friends they have), information derived from the body of Tweets and opinionatedness ratings of Tweets can improve performance. Experimental results show that social features can improve retrieval performance. Retrieval using a novel unsupervised opinionatedness feature achieves comparable performance with a supervised method using manually tagged Tweets. Topic-related specific structured Tweet sets are shown to help with query-dependent opinion retrieval. Finally, we further verify the effectiveness of our approach for opinion retrieval in re-tagged TREC Tweets2011 corpus.

Keywords Opinion retrieval · Twitter · Learning to rank · Social media

A preliminary version of this paper appears in the proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2012.

Z. Luo (✉) · T. Wang
College of Computer, National University of Defense Technology, 410073 Changsha, Hunan, China
e-mail: zhunchenluo@nudt.edu.cn

T. Wang
e-mail: tingwang@nudt.edu.cn

M. Osborne
School of Informatics, The University of Edinburgh, EH8 9AB Edinburgh, UK
e-mail: miles@inf.ed.ac.uk

1 Introduction

Twitter is a popular online social networking service which attracts over 500 million active users and generates over 340 million Tweets daily.[1] In Twitter people like to share their information or opinions about personalities, politicians, products, companies, events etc. Twitter has become an enormous repository which can not only help other people make decisions, but also help business and government collect valuable feedback. For example, Jansen et al. investigated Tweets as a form of electronic word-of-mouth for sharing consumer opinions concerning brands [17]. O'Connor et al. connected measurement of textual sentiment in Twitter and public opinion polls [31]. Bollen et al. used Twitter mood to predict the stock market [4]. However, most existing work concentrates on analysing opinions expressed in Tweets for a given topic, with none on actually finding opinionated Tweets regarding some person, product or event.

Here we present the first study of opinion retrieval in Twitter. Relevant opinionated Tweets should satisfy two criteria: (1) be relevant to the query and (2) contain opinions or comments about the query, irrespective of being positive or negative. Consider the following Tweets which are related to the topic "UK strike":

a) *"Perhaps if the public sector workers on #strike today go Christmas shopping then at least it will give the high street / UK economy a boost!"*.
b) *"UK: BBC - Up to TWO Million Set to Strike* http://t.co/wBrsgrKh *#tcot #gop #ows"*.

In these two Tweets, Tweet a) is a relevant opinionated Tweet and Tweet b) is not a relevant opinionated Tweet, since Tweet b) only describes the event without the author's opinion.

Opinion retrieval in blogs and web documents has been studied in depth [9, 11, 15, 39, 40]. The Text Retrieval Conference (TREC) recently introduced a "Blog Track" focusing on information retrieval using blog documents [25, 34, 35]. Opinion retrieval is more difficult in Twitter for the following reasons:

Short Text Tweets are short and limited to 140 characters of content. This limitation causes users to employ many abbreviations, shorten terms to compress the length of the Tweets [14]. Luo et al. found that if a Tweet contains more links, hashtags or some conventions (e.g, "@username" or "RT @username), it is likely to contain more out-of-vocabulary words [23]. Therefore, these short texts exhibit a great deal of lexical variation that only exacerbates the vocabulary mismatch problem that plagues information retrieval system [28].

Highly varied textual quality Twitter is a novel domain for opinion retrieval, with spam and wildly varying documents. While some entities such as web news sources (e.g, BBC NEWS) produce high quality texts, others create content that is personal information which contains more informal language. Since low quality texts are unlikely to yield valuable information, it is a challenge to find valuable information in Twitter.

However, Twitter also presents interesting opportunities for retrieval. The rich environment presents us with a myriad of social information over-and-above just using terms in a post (for example author information such as the number of posts) all of which potentially can improve (opinion) retrieval performance. Additionally conventions have emerged

[1] http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

in Twitter which structure Tweets and this structuring can be a valuable hint when retrieving opinionated Tweets. For example, people usually add a comment before the convention "RT @username" and many of these Tweets are likely to be subjective. Moreover the Tweets published by the BBC News (for example) often start with introductory text followed by a corresponding link. Intuitively, news media is likely to disseminate objective information. Importantly, this structural information is topic independent. This motivates us to utilize the social information and structural information for opinion retrieval in Twitter.

Here we use a standard machine learning approach to learn a ranking function for Tweets that uses the available social features and an opinionated feature in addition to traditional topic-relevant features such as BM25 (based on Okapi BM25 [36]) and VSM (based on vector space model [37]). The experimental results show that opinion retrieval performance is improved when links, mentions, author information such as the number of posts or followers and the opinionatedness of the Tweet are taken into account. Additionally we use corpus-derived method to estimate the opinionatedness value of the Tweet as a feature for ranking. However, collecting manually tagged Tweets is time-consuming. Estimating the opinionatedness of a Tweet is also a topic-dependent problem [18]. It is impossible to collect topic-related manually tagged Tweets for every topic. Therefore, we propose a novel approach, using the social information and structural information of the Tweets, to automatically generate a large number of accurate 'pseudo' subjective Tweets (PSTs) and 'pseudo' objective Tweets (POTs). These two Tweet sets can be used as a corpus to derive lexicons for estimating the opinionatedness of a new Tweet. We show that our approach can achieve comparable performance with a method which using manually tagged Tweets corpus. Our approach can also generate topic-related PSTs and POTs to a given query, which can help query-dependent opinion retrieval.

The contributions can be summarized as follows:

1) To the best of our knowledge, this is the first study of opinion retrieval in Twitter. We release all the data used in this paper,[2] which consists of 50 queries total and the corresponding relevance judgments for these queries.
2) We show that our ranking function which uses social features and an opinionatedness feature is significantly better than an optimized BM25 baseline and the VSM baseline for opinion retrieval (improving MAP by 56.82 % and 33.75 % respectively).
3) We also show that the ranking model with an opinionatedness feature, using our automatic generation of PSTs and POTs based on specific structured Tweet sets, can achieve comparable performance with a method using manually tagged Tweets. Using this method in a query-dependent scenario yields further gain.
4) Finally we test our approach with the TREC Tweets2011 which we re-tagged part of opinionated topic relevant Tweets based on original relevant Tweets [33]. The experimental results show our approach is still effective for opinion retrieval in this data.

Additionally, this study is an extension of our previous work [24]. The novelty of this work includes:

a) We test the effectiveness of our social features and opinionatedness features for opinion retrieval in Twitter based on different baselines. Comparing our approach with more baselines can better verify the effectiveness of our methods.

[2] https://sourceforge.net/projects/ortwitter/

b) We evaluate the performance of our opininatedness feature as a Tweet classifier. The experimental result shows it can classify the subjective Tweets and objective Tweets effectively.
c) We retagged TREC Tweets2011 and evaluate our approach on this new dataset for opinion retrieval in Twitter.

2 Related work

We review related work on three main areas: opinion mining and information retrieval in Twitter, opinion retrieval in TREC.

2.1 Opinion mining in Twitter

Twitter has attracted hundreds of millions of users who post opinions on this platform and it is also a hot domain for academic research.

Jansen et al. investigated Tweets as a form of electronic word-of-mouth for sharing consumer opinions concerning brands [17]. They studied the effects of services in the commercial sector, namely, the impact on the relationship between company and customer. Their research findings show that 19 % of microblogs contain mention of a brand. Of the branding microblogs, nearly 20 % contained some expression of brand sentiments. This shows Twitter has become an important platform for companies to collect customers' feedback of products.

O'Connor et al. connected measures of public opinion derived from polls with sentiment measured from analysis of Tweets [31]. They found that a relatively simple sentiment detector based on Twitter data replicates consumer confidence and presidential job approval polls. It shows expensive and time-intensive polling can be replaced by the simple-to-gather text data that is generated from Twitter.

Bollen et al. investigated whether measurements of collective mood states derived from large-scale Twitter feeds were correlated to the value of the Dow Jones Industrial Average [4]. They found that changes in the public mood state can indeed be tracked from the content of large-scale Twitter feeds. They used this changing information to predict the stock market successfully. It shows Twitter is an important source for people to collect public opinions.

All the above work concentrates on analyzing opinions expressed in Twitter for a given topic, none on how to obtain opinions towards any topic. Here we consider the problem of finding opinions given any topic in Twitter.

Another related work is sentiment analysis in Twitter. Davidov et al. used machine learning approach to identify the polarity of a Tweet [5]. They used hashtags and smileys as labels for constructing training data. Barbosa and Feng used training data collected from Twitter sentiment classification web sites (e.g., *Twendz*,[3] *Twitter Sentiment*[4] and *TweetFeel*)[5] for sentiment analysis [3]. Jiang et al. improved target-dependent sentiment analysis by incorporating target-dependent features and the context of Tweets [18]. Korenek and Šimko utilised the appraisal theory to identify sentiment that is connected to a main target of a microblog post. They created an appraisal dictionary to contain more annotated terms [20].

[3] http://twendz.waggeneredstrom.com/

[4] http://twittersentiment.appspot.com/

[5] http://www.Tweetfeel.com/

The purpose of the sentiment analysis is indicating the polarity of text, our work, however, only consider the opinionatedness of a Tweet no matter if it is positive or negative.

2.2 Information retrieval in Twitter

The large volume of real-time Tweets posted on Twitter per day are highly attractive for information retrieval research. Efron proposed a language modelling approach for hashtag retrieval [7]. He used the retrieved hashtags on a topic of interest for query expansion to improve the performance of Twitter search. Massoudi et al. studied a new retrieval model for Twitter search by considering the model with textual quality and Twitter specific quality indicators [27]. They found that this model had a significant positive impact on Tweets retrieval. Naveed et al. combined document length normalization in a retrieval model to resolve the sparsity of short texts for Tweets [30]. Duan et al. considered learning-to-rank for Tweets [6]. They proposed a new ranking strategy which uses not only the content relevance of a Tweet, but also the account authority and Tweet-specific features. Luo et al. investigated the internal structure of Tweets to improve Tweets retrieval in an ad-hoc retrieval setting [23]. The experimental results showed that the ranking approach using this structural information alone achieved comparable performance to the state-of-the-art method. Furthermore, using the structural information features together with other social media features can achieve higher performance. All the above work aims to find topic relevant documents in Twitter. However, our purpose is finding Tweets which are not only topic relevant but also are opinionated.

TREC 2011 introduced the Microblog Track which addressed one single pilot task, entitled *real-time search task*, where the user wished to see the most recent but relevant information to the query [28]. A total 59 groups participated in the track from across the world, with 184 submitted runs. The experimental results indicate the large gap between the best and medians evaluation score (e.g., MAP value) per-topic for 59 participated groups. It shows that Tweets retrieval is far from being a solved problem.

2.3 Opinion retrieval in TREC

The opinion retrieval task for blogs was firstly introduced in TREC 2006 [34] and continued in TREC 2007 and 2008 [25, 35]. Most groups participated in TREC adopt a two-stage approach, where an initial set of relevant but not necessarily opinionated documents are re-ranked by taking into account various document opinion features.

There is other work related to opinion retrieval. Eguchi and Lavrenko firstly introduced opinion ranking formula which combine sentiment relevance models and topic relevance models into a generation model [8]. This formula was shown to be effective on the MPQA corpus, but it does not perform very well in the following TREC opinion retrieval experiment. Zhang and Ye and Huang and Croft also put forward their own way to unify sentiment relevance models and topic relevance models for ranking [16, 39]. Gerani et al. firstly investigated learning-to-rank for blog posts [10]. All this work is in the context of blogs or web documents, Twitter, however, is a novel domain and its rich social environment should be considered when modeling relevance.

In opinion retrieval estimating the opinionatedness of a document is essential. He et al. proposed an approach to calculate the opinionatedness of a document based on subjective terms [15]. These terms are automatically derived from manually tagged data. They used all opinionated relevant documents to queries as a subjective document set and other topic-relevant document as an objective document set. The opinionated score of each term

can measured by the divergence of the distribution in these two sets. Amati et al. adopted a similar approach for the automatic construction of an opinion-term vocabulary for ad-hoc retrieval [2]. Seki and Uehara used a statistical language model, incorporating distant word dependencies, to model the opinionated document for ranking [38]. Jijkoun et al. present a method for automatically generating topic-specific subjective lexicons based on extracting syntactic clues of manually tagged data [19]. Li et al. proposed a novel notion of topic-sentiment word pairs as a new representation for opinion retrieval task [21]. Orimaye and Alhashmi exploited the grammatical derivation of sentences to show contextual and subjective relevance [32]. Akritidis and Alhashmi proposed an approach which integrates query-independent and time- sensitive quality metrics (QUIQS) into the current ranking schemes, and combines them with the computed relevance and opinion scores [1]. While all of the above approaches are effective for opinion retrieval, they need human tagged subjective and objective documents.

Unlike the work introduced above, Zhang et al. used the reviews of RateitAll.com and other webpages as a source of 'pseudo' subjective sentences (PSSs), and Wikipedia documents as an external source of 'pseudo' objective sentences (POSs) [40]. They assume that the subjective portion should be dominant in the reviews so that the effect of the objective portion can be neglected. The situation is opposite when using Wikipedia documents. They then used these PSSs and POSs to build an SVM sentence classifier. This classifier can give the sentence an opinionated score which is combined with the topic relevance score for ranking. An approach based on a similar idea we also proposed in the context of Twitter, which using the structural information and social information to automatically generate 'pseudo' subjective Tweets (PSTs) and 'pseudo' objective Tweets (POTs), for opinion retrieval in Twitter.

3 Overview of our approach

To generate a good ranking function for Tweets, we investigate social features and an opinionatedness feature. We develop a bag of features into a learning-to-rank scenario which demonstrated excellent power for ranking problem [22].

3.1 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. Figure 1 shows the framework of learning to rank opinions in Twitter. First, a set of queries Q 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 opinion in Twitter.

3.2 Features for Tweets ranking

For opinion retrieval in Twitter, we exploit social features and an opinionatedness feature for Tweets ranking.

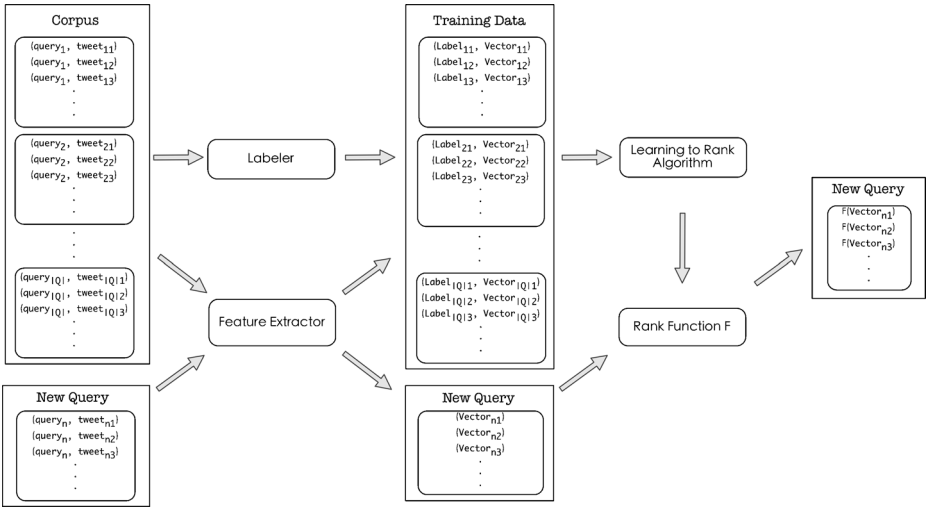

Figure 1 Framework of learning to rank opinions in Twitter

1) *Social features* refer to those features that describe the specific characteristics of Twitter and the authors' information.
2) *An Opinionatedness feature* refers to the the feature which estimate the opinionatedness of the Tweet to a given topic.

In the following sections, we describe these features in details.

4 Social features

The following features capture useful aspect of Twitter and authors for opinion retrieval.

4.1 Twitter specific features

People usually use some conventions to Tweet and these conventions might be related to the opinionatedness of the Tweets.

URL: Sharing links in Tweets is very popular in Twitter. Most Tweets containing a link usually give the objective introduction to the links (e.g., Tweets posted by the BBC News). Additionally, spam in Twitter often contain links. Hence, we use a feature indicating whether a Tweet contains a link in our ranking model. If a Tweet contains links, the URL feature score of the Tweet is 1, otherwise the score is 0.

Mention: In a Tweet, people usually use "@" preceding a user name to reply to other users. The text of this Tweet is more likely to be 'personal content'. Previous work shows that 'personal content' is on the whole more likely to contain opinions than 'official content' [12]. Therefore, we use a binary feature indicating whether a Tweet contains "@username" for Tweets opinion retrieval.

Hashtag: A hashtag refers to a word in the text of the Tweet that begins with the "#" character. It is used to indicate the topic of the Tweet. We add a binary feature whether a Tweet contains a hashtag into our system.

Recency: Twitter generates streams of text in real time and it is often hypothesized that more recent results are better for Twitter retrieval. Therefore, we use a feature that measures the age (in seconds) of the Tweet at the time the query was issued.

4.2 Author features

Twitter is a social network. The rich author information can also be used for the task.

Statuses: The number of Tweets (statuses) the author has ever written is related to the activeness of an author. Intuitively, the most active authors are likely to be spammers who post very large number of Tweets. Therefore, we use the number of statuses as a feature for Tweets ranking.

Followers and Friends: In Twitter a user can choose to follow any number of other users that he finds interesting for one reason or another. If *userA* follows *userB*, all the Tweets posted by *userB* will be updated in the *userA's* private stream. We call *userA* a *follower* of *userB* and *userB* a *friend* of *userA*. The number of followers indicates the popularity of the user. For example, the news media users usually have more followers than normal users. The number of friends also reflects the type of the user. For example, spammers often have large number of 'friends'. We develop these two features for Tweets retrieval.

Listed: A user can group their friends into different lists according to some criteria (e.g., the topic and social relationship). If a user is listed many times, it means that his Tweets are interesting to a large user population. We use a feature that measures how many times the author of a Tweet has been listed for Tweet ranking.

5 An opinionatedness feature

Obviously estimating the opinionatedness score of a Tweet is essential for an opinion retrieval task. Previous approaches for this estimation are divided into two categories: 1) classification approach and 2) lexicon-based approach [29]. We adopt the lexicon-based approach, since it is simple and not dependent on machine learning techniques. However, a lexicon such as MPQA Subjectivity Lexicon[6] which is widely used might not be effective in Twitter, since the textual content of a Tweet is often very short and highly informal. Therefore, we use a corpus-derived lexicon to construct an opinion score for each Tweet. We estimate the opinionatedness score of each Tweet by calculating the average opinion score over certain terms. We use the chi-square value, based on manually tagged subjective Tweets set and objective Tweets set, to estimate the opinion score of a term. The score measures how dependent a term is with respect to the subjective Tweets set or objective Tweets set. For all terms in a Tweet, we only keep the terms with a chi-square value no less than m when computing the opinion score. The estimated formula as follows:

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

Table 1 Table for pearson's chi-square

	t	$\neg t$	Row total
Sub. set	O_{11}	O_{12}	O_{1*}
Obj. set	O_{21}	O_{22}	O_{2*}
Col. total	O_{*1}	O_{*2}	O

$O_{1*} = O_{11} + O_{12}, O_{2*} = O_{21} + O_{22}, O_{*1} = O_{11} + O_{21}, O_{*2} = O_{12} + O_{22}, O = O_{11} + O_{12} + O_{21} + O_{22}$

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

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

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

where $sgn(*)$ is a 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 subjective/objective Tweets set respectively. For example O_{12} is the number of Tweets not having term t in the subjective Tweets set.

Manually labelling the Tweets necessary for constructing opinionated scoring is time-consuming and also topic-dependent. For example, Tweets about "android" might contain opinionated terms "*open*", "*fast*" and "*excellent*", but these terms are unlikely to be the subjective clues of Tweets related to some news events (e.g., "UK strike"). It is clearly impossible to tag a large number of Tweets for every given topic. Therefore, we develop an approach to collect 'pseudo' subjective Tweets (PSTs) and 'pseudo' objective Tweets (POTs) automatically.

In Twitter, some simple structural information of Tweets and users' information can be used to generate PSTs and POTs. For example people usually reTweet another user's Tweet and give a comment before this Tweet. Tweets with this structure are more likely to be subjective. Many Tweets posted by news agencies, who usually post many Tweets and have many followers, are likely to be objective Tweets and these Tweets usually contain links. We define these two types of Tweets as follows:

1) **'Pseudo' Subjective Tweet (PST):** a Tweet of the form "RT @username" with text before the reTweet. For example, a 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 pseudo subjective Tweet.
2) **'Pseudo' Objective Tweet (POT):** If a Tweet satisfies two criteria: (1) it contains links and (2) the user of this Tweet posted many Tweets before and has many followers. This Tweet is likely to be an objective Tweet. E.g., "*#NorthKorea:#KimJongil died after suffering massive heart attack on train on Saturday, official news agency reports* bbc.in/vzPGY5".

Using the definition introduced above, it is easy for us to construct patterns and collect a large number of PSTs and POTs from Twitter. We assume that the Tweets in the PST set are

all subjective Tweets and the Tweets in the POT set are all objective Tweets. Although this is not 100 % true, the subjective Tweets portion should be dominant in the PST set so that the effect of the objective Tweets portion can be neglected. It is opposite in the POT set. Since the structural information and authors' information are independent of the topic of a Tweet, if there are a lot of Tweets related to a given topic, it is easily to collect topic-dependent PSTs and POTs.

6 Experiments

Here section we discuss the experiment we conducted in order to evaluate the above-proposed features for opinion retrieval in Twitter.

6.1 Dataset

There exists no **ground truth** dataset for evaluating the task of opinion retrieval in Twitter. In order to identify *which of the Tweets are topic related to a query and also contain opinions about it (query)*, we therefore create a new dataset by ourselves.

We crawled and indexed about 30 million Tweets using the Twitter API in November 2011. All Tweets are in English.[7] Using these Tweets we implemented a search engine using Lucene.[8] Seven people (a woman and six men) were asked to use our search engine. All of them are not native speakers but are good at it. They were allowed to post any query. Given a query the search engine would present a list of 100 Tweets ranked based on the BM25 score. We use Lucene-BM25[9] to calculate the BM25 score of a Tweet to a query. We use default setting as the specific BM25 parameters ($k1 = 2$; $b = 0.75$). All the queries were issued on December 1, 2011.

Based on the principle about the Tweet whether it is topic related and expresses opinions about it (query), the user who issued the query assigned a binary label to every Tweet. If a Tweet is topic related to a query and also contains opinions judged by the people who issued the query, the score of Tweet class is 1, otherwise the score is 0. We emphasize that if a Tweet is only topic related or just contains opinions, it is not a relevant Tweet in our task.

Finally we collected 50 queries and 5000 judged Tweets (100 Tweets per query). We group the queries into six classes according to the domain. Table 2 shows all the queries and the number of relevant Tweets for each query. The average query length was 1.94 words and the average number of relevant (and subjective) Tweets per query was 16.62.

We considered the reliability of these relevance judgements. For each query, we randomly sampled 10 relevant Tweets (as labelled by our original judges) and asked two new annotators to decide whether they were relevant to the query in question. The two annotators had a kappa score of 0.54, which is generally considered to indicate 'good' reliability.

[7] We filtered English Tweets using toolkit language-detection from http://code.google.com/p/language-detection

[8] http://lucene.apache.org

[9] http://nlp.uned.es/~jperezi/Lucene-BM25/

Table 2 Queries and the number of relevant Tweets

Organization	Product	Person
pixar, 23	Mac book pro, 37	Jennifer Aniston, 40
manchester city, 21	iphone4s, 32	chris paul, 39
htc, 19	kinect, 30	Obama, 35
Syria, 17	itouch, 25	bill gates, 16
iran, 15	kindle fire, 16	Maggie Q, 13
Manchester United, 15	iOS5 Jailbreak, 12	owl city, 12
disney, 12	galaxy note, 11	paul graham, 10
Lenovo, 10	Xbox 360, 7	steve jobs, 9
microsoft, 6	google venture, 7	Kai-Fu Lee, 1
Calvin Klein, 5	new Google Bar, 5	
fossil, 5	EA Daily Deals, 0	
intel, 3	SIEMENS fridge freezer, 0	
channel, 0		

Others	Event	Movie and TV Series
job hunting, 79	iran nuclear, 35	Breaking Dawn, 49
speech recognition, 15	American Music Awards, 25	big bang, 41
machine learning, 4	UK embassy, 21	Two And A Half Men, 35
new start-ups, 2	UK strike, 7	inside job, 2
immigrate to canada, 1	ARTIST OF THE YEAR, 7	
systems biology, 0		
text mining, 0		

6.2 Experimental settings

We investigate the effect of features introduced above for opinion retrieval in Twitter. For learning to rank, SVM^{Rank} which implements the ranking algorithm is used.[10] 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 5-fold cross-validation in our dataset. We split out data into training, testing and validation data. The splits are within folds. We choose Tweets of 30 queries as the training data. The remaining Tweets are divided into testing data and validation data equally. There are several metrics that can be used to measure the performance of ranking. Here, we use *Mean Average Precision* (MAP) as the evaluation metric, since it is most standard among the TREC community which has been shown to have especially good discrimination and stability [26].

6.3 Baselines

Comparing our approach with more baselines can better verify the effectiveness of our method for opinion retrieval in Twitter. Therefore, in our experiments, we used two different baselines: One is **BM25** with parameters optimized for Twitter which uses the Okapi BM25

[10] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

score of each Tweet as a feature for modeling; the other baseline is **VSM** which we use the vector space model to calculate the content relevance of the query and the Tweet [37].

In information retrieval, Okapi BM25 is the best of the known probabilistic weighting schemes which ranks matching documents according to their relevance to a given search query. Given a query q, containing term t, the BM25 score of a document d is:

$$BM25(q,d) = \sum_{t \in q} log\left(\frac{N - df_i + 0.5}{df_i + 0.5} * \frac{(k_1+1)(tf_i)}{k_1\left((1-b) + b\frac{dl}{avdl}\right) + tf_i}\right)$$

Where tf_i represents the document term frequency, N is the total number of documents in the collection, dl is the document length and $avdl$ is the average document length in the collection. The two parameters k_1 and b control the influence of term frequency and adjust the document length normalization respectively. We use the validation data in each fold to optimize the parameters $k1$ and b for Tweets ranking when calculates the Okapi BM25 score for each Tweet.

Since Tweets are short texts and BM25 is well known for evening out high variation in the length of documents without fully normalizing for length, **BM25** might not be the most

Table 3 The summarization of features

Baseline features	Value range	Description
BM25	$(0, +\infty)$	The BM25 score
VSM	$(0, +\infty)$	The VSM score
Social media features	**Value range**	**Description**
URL	0 or 1	Whether the Tweet has links
Mentions	0 or 1	Whether the Tweet has mentions
Hashtags	0 or 1	Whether the Tweet has hashtags
Recency	$N^+ = \{1, 2, 3, ...\}$	How long (in seconds) did the user publish the Tweet before the query submitted
Followers	$N = \{0, 1, 2, ...\}$	The number of followers the author has
Friends	$N = \{0, 1, 2, ...\}$	The number of friends the author has
Listed	$N = \{0, 1, 2, ...\}$	The number of lists the author appears in
Statuses	$N^+ = \{1, 2, 3, ...\}$	The number of statuses the author has
Opinionatedness features	**Value range**	**Description**
MPQA_Lexicon	0 or 1	Whether the Tweet has a word or a phrase in MPQA Subjectivty Lexicon
TwitterSenti	0 or 1	The opinionatedness score judged by Twitter Sentiment API
Gold	$(-\infty, +\infty)$	The opinionatedness score estimated by manually tagged Tweets
Q_I	$(-\infty, +\infty)$	The opinionatedness score estimated by topic-independent PSTs and POTs
Q_D	$(-\infty, +\infty)$	The opinionatedness score estimated by topic-dependent PSTs and POTs

Table 4 Performance of ranking method using social features based on BM25

	MAP
BM25	0.2831
BM25+URL	0.3305▲
BM25+Mention	0.2920
BM25+Hashtag	0.2734
BM25+Recency	0.2576
BM25+Statuses	0.2931
BM25+Followers	0.2946△
BM25+Friends	0.2799
BM25+Listed	0.2822

A significant improvement over the BM25 ranking method with △ and ▲ (for $p < 0.05$ and $p < 0.01$)

natural baseline. We use the classic **VSM** as the other baseline proposed by Salton et al. [37]. In VSM, a documents d_j and a query q can be represented as t-dimension vectors $d_j = (w_{1,j}, w_{2,j}, ..., , w_{t,j})$ and $q = (w_{1,q}, w_{2,q}, ..., , w_{t,q})$, where t is the number of indexed terms and each dimension corresponds to a separate term. There are several different ways of computing term weight. We use the best known schemes $tf\text{-}df$ weighting, where

$$w_{t,d} = tf_{t,d} * log\left(\frac{|D|}{|d' \in D|t \in d'|}\right)$$

and $tf_{t,d}$ is term frequency of term t in document d, $log\left(\frac{|D|}{|d' \in D|t \in d'|}\right)$ is inverse document frequency. $|D|$ is the total number of documents in the document set; $|d' \in D|t \in d'|$ is the number of documents containing the term t. The similarities between d_j and q can be calculated by comparing the deviation of angles (cosine similarity) between the document vector and the query vector:

$$sim(d_j, q) = \frac{\sum_{i=1}^{N} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} * \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

In our experiments, we take advantage of Lucene to implement VSM.

Table 5 Performance of ranking method using social features based on VSM

	MAP
VSM	0.2812
VSM+URL	0.3171▲
VSM+Mention	0.2932△
VSM+Hashtag	0.2803
VSM+Recency	0.2757
VSM+Statuses	0.2928△
VSM+Followers	0.2829
VSM+Friends	0.2808
VSM+Listed	0.2801

A significant improvement over the VSM ranking method with △ and ▲ (for $p < 0.05$ and $p < 0.01$)

Table 6 Performance of ranking method using different opinionatedness features based on BM25

	MAP
BM25	0.2831
BM25+MPQA_Lexicon	0.2895
BM25+TwitterSenti	0.3279▲
BM25+Gold	0.3739▲
BM25+Q_I	0.3792▲

A significant improvement over the BM25 ranking method with △ and ▲ (for $p < 0.05$ and $p < 0.01$)

6.4 Results

We structure our main results as follows:

- Can social features improve opinion retrieval in Twitter estimated by our dataset? (Section 6.4.1)
- Can opinionatedness features improve opinion retrieval in Twitter? (Section 6.4.2)
- Do topic-dependent PSTs and POTs help beyond topic-independent PSTs and POTs for opinion retrieval in Twitter? (Section 6.4.3)
- Do PSTs and POTs help classify subjective Tweets? (Section 6.4.4)
- Which is the best ranking model for Twitter opinion retrieval? (Section 6.4.5)

Table 3 summarizes the features we used.

6.4.1 Social features evaluation

We first investigate whether social features can improve opinion retrieval in Twitter. We combine each social feature with the baselines within our Tweet ranking systems. Tables 4 and 5 shows the performance of each ranking model. We can see that **URL** and **Followers** features can significantly improve the results when used with the baseline **BM25** in isolation in Table 4.[11] Although **Mention** and **Statuses** features improve the results, they are not significant. In Table 5, we can see the **URL**, **Mention** and **Statuses** features can improve this task significantly. All these suggest some social information can indeed help the opinion retrieval in Twitter. Especially, the **URL** feature is the most effective feature, perhaps because most textual content in these Tweets are objective introductions. Also, spammers usually post Tweets including links and the feature dealing the links might help reduce spam. The effect of the **URL**, **Statuses** and **Followers** features for Tweets ranking also supports our approach of using social information and structural information to generate 'pseudo' objective Tweets. The improvement of ranking result using the **Mention** feature supports the idea that "personal content" is on the whole more likely to contain opinions than "official content" [12]. Somewhat surprisingly, the **Recency** feature is not effective for opinion retrieval in Twitter as we expected. We believe the reason for this is that all the queries were issued when the whole one month dataset had been collected and most of queries are not new events (see Table 2), therefore the freshness of information is not very important factor for the requirement of our users.

[11] We use t-test for statistical significance test.

Table 7 Performance of ranking method using different opinionatedness features based on VSM

	MAP
VSM	0.2812
VSM+MPQA_Lexicon	0.2876
VSM+TwitterSenti	0.3244▲
VSM+Gold	0.3485▲
VSM+Q_I	0.3566▲

A significant improvement over the VSM ranking method with △ and ▲ (for $p < 0.05$ and $p < 0.01$)

6.4.2 Opinionatedness feature evaluation

Next we investigate the opinionatedness feature for Tweets ranking. To automatically generate 'pseudo' subjective Tweets (PSTs) and 'pseudo' objective Tweets (POTs), we design some simple patterns: for PSTs generation, we choose the Tweets which use the convention "RT @username", with the text before the first occurrence[12] of this convention. Additionally we find 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 PSTs and 1.35 % Tweets are POTs.

We ask one person (non-authors) to spot-checked the quality of our automatically harvested Tweets by randomly selected 100 PSTs and 100 POTs and manually inspected them, judging the extent to which there were subjective or objective. In these Tweets, 95 % PSTs were subjective Tweets and 85 % POTs were objective Tweets. This supports the idea that our approach can generate a large number of accurate PSTs and POTs. Hence, we randomly choose 3000 English PSTs and POTs to form a topic-independent dataset.

In our corpus-derived approach, we use the Porter English stemmer and stop words[13] 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 to 5.02 corresponding to the significance level of 0.025 for each term in dataset. This setting is the same as [40]'s work. We call the feature using topic-independent dataset to estimate the opinionatedness score **Q_I**. Previous work uses manually tagged blogs to estimate the opinionatedness score of a new blog [10, 15]. In our experiment we use the manually tagged Tweets of training data in each fold to estimate the opinionatedness score of a new Tweet. We compare the method, using **Gold** feature based on these manually tagged Tweets, with the method using our **Q_I** feature for Tweets ranking. We also develop another two features for comparison. **MPQA_Lexicon** feature: if a Tweet contains a word or a phrase in MPQA Subjectivity Lexicon, the opinionatedness score of the Tweet is 1, otherwise the score is 0. **TwitterSenti** feature: If the Tweet is a positive Tweet or a negative Tweet judged by public Twitter Sentiment API[14] [13], the Tweet's opinionatedness score is 1. This API judges the Tweet as a neutral Tweet, the Tweet's opinionatedness score is 0.[15]

[12] The Tweet might contain more than one "RT @usename.

[13] It contains standard stop words, commonly used punctuation and the Twitter convention "RT".

[14] http://www.sentiment140.com

[15] It does not mean the neutral Tweet is an objective Tweet. Actually some subjective Tweets have no polarity, but it is out of consideration in this paper.

Table 8 Performance of ranking method using Q_D feature based on BM25

	MAP
BM25+Q_I	0.3792
BM25+Q_D	0.3907$^{\triangle}$

A significant improvement over the BM25+Q_I ranking method with $^{\triangle}$ and $^{\blacktriangle}$ (for $p < 0.05$ and $p < 0.01$)

Tables 6 and 7 show the results of ranking using the opinionatedness features. We can see that all the methods using opinionatedness features improve the opinion retrieval performance over the baselines. It shows estimating the opinionatedness score of a Tweet is essential for opinion retrieval task. Although the BM25+MPQA_Lexicon method and VSM+MPQA_Lexicon method, using the **MPQA_Lexicon** feature, can improve Tweet ranking, the results achieved are worse than the other ranking methods. The reason is that the text of Tweets is different from other documents (e.g, reviews and blogs) and the MPQA Subjectivity Lexicon is not effective enough for Twitter analysis. The ranking method using **Q_I** feature can achieve comparable performance with the BM25+Gold method (there are no significant difference at p=0.05). It suggests that using social information and structural information to generate accurate PSTs and POTs automatically is useful for opinion retrieval in Twitter. Importantly this method does not need any manually tagged Tweets.

6.4.3 Feature based on topic-dependent PSTs and POTs evaluation

Another advantage of our approach is that it is easy to gather topic-dependent PSTs and POTs. We use all PSTs and POTs introduced above to implement a search engine. Given a query, the search engine can give any number of query-dependent PSTs and POTs ranked by topic relevance. We generate 3000 query-dependent PSTs and POTs for each query. Using each query-dependent Tweets we calculate the opinionatedness feature (called **Q_D** feature) for a new Tweet. Tables 8 and 9 show the result of ranking methods using the **Q_D** feature. It improves the opinion retrieval in Twitter over the BM25+Q_I ranking method and VSM+Q_I ranking method which do not consider query-dependent situation. It means our approach can help resolving query-dependent problem for opinion retrieval in Twitter.

Table 10 shows a list of the highest score of $\chi^2(t)$ opinion terms derived from different query-dependent PSTs and POTs and query-independent PSTs and POTs (totally 1000 Tweets for each query). We can see that our approach can assign high scores to terms such as personal pronoun (e.g., "*i*, "*u*" and "*my*") and emoticons (e.g., "*:)*", "*:(*" and "*:d*"). The reason is that personal content Tweets are more likely to be subjective Tweets. For query-dependent PSTs and POTs, our approach successfully extracts the opinionated feature "*excit*" ($Opinion(t) > 0$) which can express attitude about the movie "Breaking Dawn", and this term is unlikely to be used in the opinionated Tweets related to "UK strike" topic.

Table 9 Performance of ranking method using Q_D feature based on VSM

	MAP
VSM+Q_I	0.3566
VSM+Q_D	0.3599

A significant improvement over the VSM+Q_I ranking method with $^{\triangle}$ and $^{\blacktriangle}$ (for $p < 0.05$ and $p < 0.01$)

Table 10 The highest score of $\chi^2(t)$ opinion terms derived from different query-dependent PSTs and POTs and query-independent PSTs and POTs

Rank	Breaking Dawn	HTC	Obama	UK strike	Q_I
1	i +	i +	i +	... -	i +
2	video -	lol +	you +	i +	lol +
3	go +	.. +	#obama -	followfridai -	:) +
4	.. +	u +	my +	rank -	.. +
5	me +	my +	lol +	you +	u +
6	lol +	new -	u +	my +	* +
7	new -	:) +	!! +	lol +	new -
8	via -	me +	me +	week -	my +
9	!!! +	* +	barack -	last -	morn +
10	wait +	rezound -	#tcot -	:) +	me +
11	pattinson -	you +	... -	u +	!!! +
12	robert -	phone -	cont +	me +	good +
13	... -	like +	.. +	thi +	:d +
14	so +	:d +	presid -	so +	via -
15	too +	!!! +	* +	!! +	!! +
16	:) +	morn +	i'm +	#ows -	cont +
17	see +	i'm +	:) +	#jobs -	haha +
18	can't +	good +	we +	x +	ya +
19	:d +	!! +	do +	come +	too +
20	premier -	... -	he +	3 +	... -
21	kristen -	too +	obama' -	gener -	i'm +
22	excit +	cream -	!!! +	onli +	:(+
23	again +	cont +	#news -	good +	thank +
24	i'm +	so +	know +	bbc -	,) +
25	im +	thank +	lmao +	here +	@damnitstrue +

"+" is the score of $Opinion(t)$ no less than 5.02. "−" is the score of $Opinion(t)$ no more than -5.02

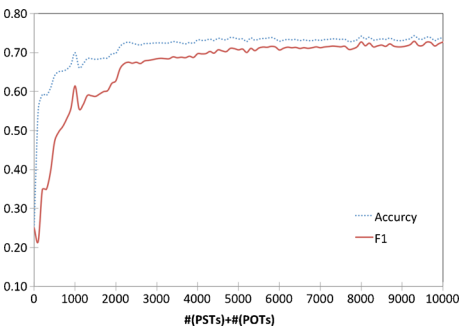

Figure 2 Performance of Tweets Classifier based on Opinion_avg(d)

Table 11 Performance of the best ranking method based on BM25

	MAP
BM25	0.2831
BM25_Best	0.4181▲
BM25_All	0.4128▲

A significant improvement over the BM25 ranking method with △ and ▲ (for $p < 0.05$ and $p < 0.01$)

In PSTs and POTs related to the "UK strike" topic, we discover (unsurprisingly) that the term "*bbc*" ($Opinion(t) < 0$) is more likely to appear in the objective Tweets posted by BBC news.

6.4.4 Subjective Tweets classification

We are also interested in how the **Opinion_avg(d)** formula (see Section 5) would perform if used as a Tweet classifier. We use 1000 manually tagged subjective Tweets and 1000 manually tagged objective Tweets as a gold testing data. We take a Tweet d which the value of $Opinion\_avg(d)$ is more than 0 as a subjective Tweet and a Tweet which the value is no more than 0 as an objective Tweet. Both accuracy and F1 are our classification evaluation metric. Figure 2 shows the performance of **Opinion_avg(d)** formula as a Tweet classifier with different numbers of PSTs and POTs. We can see that the performance of classifier using more than approximately 2200 PSTs and POTs can achieve the value of accuracy 0.72 and F1 0.67. There is no significant improvement when using more PSTs and POTs. All these shows using the **Opinion_avg(d)** formula as a classifier can judge the subjective Tweets effectively.

6.4.5 All features evaluation

Finally we add all the features which can improve the opinion retrieval in Twitter into a ranking model. They are **BM25** (**VSM**), **URL**, **Mention**, **Statuses**, **Followers** and **Q_D** features. Tables 11 and 12 show the best results (**BM25_Best** and **VSM_Best**) of our methods which improve MAP by 56.82 % over the BM25 ranking method and 33.75 % over the VSM ranking method. Additionally we add all features (see Table 3) into a ranking model, [16] which are called **BM25_All** and **VSM_All** respectively. Tables 11 and 12 show the performance of two ranking models are both better than two baselines. They are slightly worse than the two best ranking models, but there are no significant difference. All these shows only using **BM25** (**VSM**), **URL**, **Mention**, **Statuses**, **Followers** and **Q_D** features can achieve the best performance for opinion retrieval in Twitter.

6.5 Dataset bias discussion

Since there is no annotated dataset for opinion retrieval in Twitter, we use a dataset constructed by ourselves. However, judgement level of about 100 results per query and only using one run BM25 to collect judgments may cause the collection to be non-reusable for

[16] Here we only use the feature **Q_D** to estimate the opinionatedness score of a Tweet, since it has the best performance than the other opinionatedness features

Table 12 Performance of the best ranking method based on VSM

	MAP
VSM	0.2812
VSM_Best	0.3761▲
VSM_All	0.3721▲

A significant improvement over the VSM ranking method with △ and ▲ (for $p < 0.05$ and $p < 0.01$)

evaluating other algorithms, and a bias toward BM25-like algorithms may exhibit. Basically, when only using top results from BM25 to provide judgments, all other retrieval algorithms evaluated will show a lower bound score. That is because all unjudged documents will be seen as irrelevant. That's why TREC uses many more than 2 different retrieval algorithms to pool results and evaluates to the depth of 1000 or more through sampling. For these reasons, we use the TREC Tweets2011 set of Tweets and relevance judgements to estimate the effectiveness of our approach for opinion retrieval in Twitter.[17]

The TREC Tweets2011 corpus is comprised of 16M Tweets spread over two weeks in 2011, sampled courtesy of Twitter [33]. A total of 59 groups participated in the track from across the world and were encouraged to rank their submitted runs by preference. For the evaluation of TREC 2011 participating systems, 49 topics were created. Every group was asked to submit 30 Tweets to each queries. From the pool of 50,324 Tweets formed from the participants runs for these topics, 2,965 were judged relevant.[18] Unfortunately, the TREC Tweets2011 corpus only examines search tasks and evaluation methods for searching topic relevant Tweets, but our purpose is finding opinionated Tweets in Twitter. Therefore, we manually tagged some parts of Tweets in TREC Tweets2011 corpus which are relevant to the query and contain opinions or comments about the query. We tagged 1470 Tweets which submitted by [28] in the Microblog track (the Run ID is isiFDL), since their submitted Tweets set achieves the best result for TREC-2011 microblog track [33] and it contains the most topic relevant Tweets. Finally, there are 98 Tweets which are topic relevant and contain opinions.

Our approach is originally designed for opinion retrieval. It is interesting to know how effective this approach is when it is applied to re-rank the new tagged TREC dataset. Here, we used the baseline based on [28] which makes used of best-practice ranking techniques for searching topic relevant Tweets. We call this baseline **isiFDL**. Then we add the **URL**, **Mention**, **Statuses**, **Followers** and **Q_D** features, which can improve the opinion retrieval in Twitter, into a ranking model (**isiFDL_Best**). We also perform 5-fold cross-validation in this dataset. Table 13 shows the result. We can see that the performance of **isiFDL_Best** is significantly better than **isiFDL**. It means that our approach is still effective when using the new tagged TREC Tweets2011 corpus.

Obviously, although this new dataset minimizes the bias in topic relevance, it may still be present bias in opinionatedness relevance. It is infeasible for us to construct the testing collections for opinion retrieval in Twitter like TREC style evaluation. In the future, if there is a Twitter dataset for opinion retrieval like TREC style, we will examine our approach further.

[17] http://trec.nist.gov/data/Tweets/

[18] http://trec.nist.gov/data/microblog/11/microblog11-qrels

Table 13 Performance of our opinion retrieval approach in TREC Tweets2011 data

	MAP
isiFDL	0.1639
isiFDL_Best	0.2181$^{\triangle}$

A significant improvement over the isiFDL ranking method with $^{\triangle}$ and $^{\blacktriangle}$ (for $p < 0.05$ and $p < 0.01$)

7 Conclusion

To the best of our knowledge, we are the first to propose a ranking model for opinion retrieval in Twitter. This model integrates social and opinionatedness information for Tweets opinion retrieval. The experimental results show that opinion retrieval performance is improved when links, mentions, author information such as the number of statues or followers and the opinionatedness of the Tweet are taken into account. We also propose a novel approach which uses the social information and structural information of the Tweets to generate accurate 'pseudo' subjective Tweets (PSTs) and 'pseudo' objective Tweets (POTs) automatically. Opinionated retrieval results using this information is comparable to results using manually labelled data. When using this approach considering query-dependent situation can yield high performance.

Acknowledgments We would like to thank Victor Lavrenko for his comments and Bo Dai, Ming Lei, Zhengshuai Lin, Xianling Mao, Hongguang Ren, He Wang, Chenkun Wu and Xianglilan Zhang for tagging the data. The research is supported by the National Natural Science Foundation of China (Grant No. 61170156, 61202337 and 60933005) and a CSC scholarship.

References

1. Akritidis, L., Bozanis, P.: Improving opinionated blog retrieval effectiveness with quality measures and temporal features. World Wide Web, 1–22 (2013)
2. Amati, G., Ambrosi, E., Bianchi, M., Gaibisso, C., Gambosi, G.: Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In: Proceedings of the IR research, 30th European conference on Advances in information retrieval, ECIR'08, pp. 89–100. Springer-Verlag, Berlin, Heidelberg (2008)
3. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics, Posters, COLING '10, pp. 36–44. Association for Computational Linguistics, Stroudsburg (2010)
4. Bollen, J., Mao, H., Zeng, X.J.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
5. Davidov, D., Tsur, O., Rappoport, A.: Enhanced sentiment learning using twitter hashtags and smileys. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pp. 241–249. Association for Computational Linguistics, Stroudsburg (2010)
6. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.: An empirical study on learning to rank of tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 295–303. Association for Computational Linguistics (2010)
7. Efron, M.: Hashtag retrieval in a microblogging environment. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 787–788. ACM (2010)
8. Eguchi, K., Lavrenko, V.: Sentiment retrieval using generative models. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06, pp. 345–354. Association for Computational Linguistics, Stroudsburg (2006)
9. Gerani, S., Carman, M., Crestani, F.: Aggregation methods for proximity-based opinion retrieval. ACM Trans. Inf. Syst. (TOIS) **30**(4), 26 (2012)

10. Gerani, S., Carman, M.J., Crestani, F.: Investigating learning approaches for blog post opinion retrieval. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, ECIR '09, pp. 313–324. Springer-Verlag, Berlin, Heidelberg (2009) doi:10.1007/978-3-642-00958-7_29
11. Gerani, S., Carman, M.J., Crestani, F.: Proximity-based opinion retrieval. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 403–410. ACM (2010)
12. Gerani, S., Keikha, M., Carman, M., Crestani, F.: Personal blog retrieval using opinion features. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval, ECIR'11, pp. 747–750. Springer-Verlag, Berlin, Heidelberg (2011)
13. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Processing, pp. 1–6 (2009)
14. Gouws, S., Metzler, D., Cai, C., Hovy, E.: Contextual bearing on linguistic variation in social media. In: Proceedings of the Workshop on Languages in Social Media, LSM '11, pp. 20–29. Association for Computational Linguistics, Stroudsburg (2011)
15. He, B., Macdonald, C., He, J., Ounis, I.: An effective statistical approach to blog post opinion retrieval. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08. pp. 1063–1072. ACM, New York (2008)
16. Huang, X., Croft, W.B.: A unified relevance model for opinion retrieval. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, pp. 947–956. ACM, New York (2009)
17. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth, pp. 2169–2188. Wiley, New York. (2009) doi:10.1002/asi.v60:11
18. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT '11, vol. 1, pp. 151–160. Association for Computational Linguistics, Stroudsburg (2011)
19. Jijkoun, V., de Rijke, M., Weerkamp, W.: Generating focused topic-specific sentiment lexicons. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pp. 585–594. Association for Computational Linguistics, Stroudsburg (2010)
20. Korenek, P., Šimko, M.: Sentiment analysis on microblog utilizing appraisal theory. World Wide Web, pp. 1–21 (2013)
21. Li, B., Zhou, L., Feng, S., Wong, K.F.: A unified graph model for sentence-based opinion retrieval. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pp. 1367–1375. Association for Computational Linguistics, Stroudsburg (2010)
22. Liu, T.Y.: Learning to rank for information retrieval. Found. Trends Inf. Retr. **3**(3), 225–331 (2009). doi:10.1561/1500000016
23. Luo, Z., Osborne, M., Petrovic, S., Wang, T.: Improving twitter retrieval by exploiting structural information. In: AAAI '12: Proceedings of the Twenty-Sixth AAAI (2012)
24. Luo, Z., Osborne, M., Wang, T.: Opinion retrieval in twitter. In: 6th International AAAI Conference on Weblogs and Social Media (2012)
25. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec 2007 blog track. In: TREC (2007)
26. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
27. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. Advances in Information Retrieval, pp. 362–367 (2011)
28. Metzler, D., Cai, C.: Usc/isi at trec 2011: Microblog track. In: TREC (2011)
29. Na, S.H., Lee, Y., Nam, S.H., Lee, J.H.: Improving opinion retrieval based on query-specific sentiment lexicon. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, ECIR '09, pp. 734–738. Springer-Verlag, Berlin, Heidelberg (2009). doi:10.1007/978-3-642-00958-7_76
30. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.: Searching microblogs: coping with sparsity and document quality. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 183–188. ACM (2011)
31. 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)
32. Orimaye, S.O., Alhashmi, S.M., Siew, E.G.: Can predicate-argument structures be used for contextual opinion retrieval from blogs? World Wide Web, pp. 1–29 (2012)
33. Ounis, I., Lin, J., Soboroff, I.: Overview of the trec-2011 microblog track. In: TREC (2011)
34. Ounis, I., Macdonald, C., de Rijke, M., Mishne, G., Soboroff, I.: Overview of the trec 2006 blog track. In: TREC (2006)

35. Ounis, I., Macdonald, C., Soboroff, I.: Overview of the trec 2008 blog track. In: TREC (2008)
36. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M., et al.: Okapi at trec-3. NIST SPECIAL PUBLICATION SP, pp. 109–109 (1995)
37. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
38. Seki, K., Uehara, K.: Adaptive subjective triggers for opinionated document retrieval. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, WSDM '09, pp. 25–33. ACM, New York (2009). doi:10.1145/1498759.1498805
39. Zhang, M., Ye, X.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, pp. 411–418. ACM, New York (2008). doi:10.1145/1390334.1390405
40. Zhang, W., Yu, C., Meng, W.: Opinion retrieval from blogs. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM '07, pp. 831–840. ACM, New York (2007). doi:10.1145/1321440.1321555