Journal of Universal Computer Science, vol. 22, no. 5 (2016), 608-629
submitted: 14/9/15, accepted: 29/4/16, appeared: 1/5/16 © J.UCS

Opinion Retrieval for Twitter Using Extrinsic Information

Yoon-Sung Kim
(Korea University, Seoul, Republic of Korea
yskim@nlp.korea.ac.kr)

Young-In Song
(Korea University, Seoul, Republic of Korea
youngin.song@gmail.com)

Hae-Chang Rim[1]
(Korea University, Seoul, Republic of Korea
rim@nlp.korea.ac.kr)

Abstract: Opinion retrieval in social networks is a very useful field for industry because it can provide a facility for monitoring opinions about a product, person or issue in real time. An opinion retrieval system generally retrieves topically relevant and subjective documents based on topical relevance and a degree of subjectivity. Previous studies on opinion retrieval only considered the intrinsic features of original tweet documents and thus suffer from the data sparseness problem. In this paper, we propose a method of utilizing the extrinsic information of the original tweet and solving the data sparseness problem. We have found useful extrinsic features of related tweets, which can properly measure the degree of subjectivity of the original tweet. When we performed an opinion retrieval experiment including proposed extrinsic features within a learning-to-rank framework, the proposed model significantly outperformed both the baseline system and the state-of-the-art opinion retrieval system in terms of Mean Average Precision (MAP) and Precision@K (P@K) metrics.

Keywords: Opinion Retrieval, Sentiment Analysis, Opinion Mining, Social Media
Categories: H.3.1, H.3.3, H.4.1

1 Introduction

Twitter is one of the most popular social network services in the world. As many people use tweet to express their opinions and the amount of tweets becomes getting larger, Twitter has become an important medium that can capture the various opinions of people. There are many studies analyzing tweet messages such as sentiment analysis and opinion mining (e.g. [Go et al. 2009, Barbosa and Feng 2010, Saif et al. 2014, Tang et al. 2015, Severyn et al. 2015]).

Opinion retrieval is to retrieve relevant documents that have subjective statements toward the given query [Lee et al. 2012]. If a good opinion retrieval system is developed for retrieving relevant and subjective documents about a topic, it can be

[1] Corresponding author

utilized for figuring out the author's opinions and further establishing a marketing strategy reflecting author's opinions.

Previous opinion retrieval systems were developed mainly for retrieving topically relevant and subjective blogs. However, these systems will suffer from the data sparseness problem when they try to retrieve tweets because of their short length. The firstly proposed opinion retrieval system targeting Twitter [Luo et al. 2012] tried to solve this problem by utilizing Twitter-specific features. Their work used author related meta information only for filtering spam authors, and the presence or absence of hashtags and links were utilized as Twitter-specific features. However, they did not fully utilize the Twitter-specific intrinsic and extrinsic features, and thus still suffered from the data sparseness problem.

The proposed opinion retrieval system ranks relevant and subjective tweets based on the learning-to-rank framework, which is appropriate for considering various kinds of features. In developing the system, we observed that we can classify subjective authors and objective authors based on the tweets they write, and that we can classify subjective tweets and objective tweets among hashtag related tweets based on word usage in hashtags and tweets, and that the arrangement of words and the syntactic structure of reply tweets are very different between subjective tweets and objective tweets. These observations suggest that author related tweets, hashtag related tweets and reply related tweets are good sources for estimating the subjectivity of the original tweet. Furthermore, we introduce specific intrinsic and extrinsic features we have to extract from author related tweets, hashtag related tweets and reply related tweets. This paper will discuss how the proposed features alleviate the data sparseness problem and improve the performance of the state-of-the-art system. We can summarize contributions in this paper as follows:

- We propose a method of utilizing both intrinsic and extrinsic features of a tweet when we estimate the degree of subjectivity which is a primarily important task in opinion retrieval.
- We propose useful clues and related features that can be used for estimating the subjectivity of tweets alleviating the data sparseness problem.
- We show that the proposed opinion retrieval system based on a learning-to-rank framework outperforms state-of-the-art systems.
- We construct a good-quality opinion retrieval test collection with the aid of the crowdsourcing method. The data can be released only for research purposes.

The rest of this paper is organized as follows. We review related works in section 2, and present new observations in section 3. Our proposed opinion retrieval system is presented in section 4 along with a description of its features. Finally, section 5 presents experiments and results, and section 6 concludes the paper.

2 Related Works

2.1 Opinion Retrieval in Blog

Since 2006, the Text Retrieval Conference (TREC) has continued blog track and many researchers have participated in this task ([Ounis et al. 2006, Macdonald et al. 2007, Ounis et al. 2008]). Many opinion retrieval systems are developed by using the TREC's blog corpus.

[Eguchi and Lavrenko 2006] firstly proposed a sentiment information retrieval models within the framework of probabilistic language models. [Zhang and Meng 2007] also proposed an opinion retrieval system which was developed based on topical relevance and subjectivity by using the blog data of TREC 2006. In this work, topical relevance denotes to relevance of a document with respect to a given query, and subjectivity denotes the strength of opinion the document has. They proposed opinion retrieval models using a learning-to-rank framework by estimating these two factors. [Zhang and Ye 2008] and [Huang and Croft 2009] also used TREC's blog track data and proposed generative ranking models considering topical relevance and subjectivity.

[Gerani et al. 2011] and [Lee et al. 2012] proposed opinion retrieval models which used the similar generative models of [Zhang and Ye 2008] and [Huang and Croft 2009] but additionally included opinion-relatedness by measuring the distance between a topical word and an opinionated word. However, opinion-relatedness cannot be used when we develop an opinion retrieval system targeting tweets because of their short length.

2.2 Ad-hoc Retrieval in Twitter

The Text REtrieval Conference(TREC) started the microblog track in 2011, and many participants have proposed their tweet retrieval systems in the microblog track. ([Ounis et al. 2011], [Soboroff et al. 2012], [El-ganiany et al. 2013] ,and [Lin et al. 2014]). They have tried to develop a system that can retrieve relevant tweets in real time against a given query. The retrieval system proposed by [McCredie and Macdonald 2013] has improved performance by utilizing hyperlinks connected among tweet documents. The system proposed by [El-Ganiany et al. 2014] has also improved retrieval performance by applying pseudo relevance feedback using the expanded information of tweet document links. The system proposed by [Onal et al. 2015] applied a word embedding technique for retrieving tweets, and [Rao et al. 2015] developed a tweet retrieval system that can combine lexical feedback and temporal feedback.

2.3 Opinion Retrieval in Twitter

[Luo et al. 2012] first introduced an opinion retrieval model targeting tweets. The proposed opinion retrieval model used similar subjective features to those in [Zhang and Meng 2007] and tried to utilize other features extracted from Twitter's author meta information. However, the features used are not sufficient for solving the data sparseness problem.

[Atkinson et al. 2015] proposed an opinion retrieval system using Twitter's reply threads. They identified named entity related features extracted from reply threads, and utilized those features to improve the performance of the opinion retrieval system. However, the small amount of reply threads contribute little to the estimation of topical relevance, therefore does not contribute to alleviating data sparseness in subjectivity estimation.

There are several applications of opinion retrieval systems. [Luo et al. 2013] used an opinion retrieval system to estimate the propagation power of a tweet. [Zhang et al. 2013] developed a personalized opinion retrieval system reflecting personal emotional states.

3 New Observations for Estimating the Subjectivity of Tweets

As described above, the data sparseness problem is the major obstacle to estimate the subjectivity of tweets. In this section, we introduce some clues that can be used for providing extra sources for subjectivity estimation and useful extrinsic features that can be easily extracted from related tweets.

In this preliminary study, we observe that the author related tweets, hashtag related tweets, and the reply list of tweets are good sources for estimating the subjectivity of the original tweet. Figure 1 consists of an original tweet (i.e. *Tweet #1*), three different kinds of clues (i.e. author, hashtag, and reply), and related tweets.

Figure 1: Original tweet, clues, and related tweets

As shown in Figure 1, *Tweet #1* is linked to a set of tweets written by the same author *VictoriaJustice,* and to two sets of tweets associated with two hashtags *#2TrendsIn1Day* and *#socray*. Three replies to the original tweet are also constructed into the reply tweet collection. We analyze three different outside collections of tweets and identify useful features represented in outside tweet collections for subjectivity estimation.

First of all, we observe that we can classify the subjective authors and the objective authors based on the tweets they wrote. Figure 2 shows an example of subjective tweets and objective tweets.

Subjective Tweets	Objective Tweets
I liked an @YouTube video http://t.co/Px5iXpfE Kelly Clarkson - Already Gone Live at AMA(American Music Awards	John Evans Atta Mills, the president of Ghana, has died, his chief of staff says. http://t.co/vnFE2rBP
come on htc i love htc	#Celebrity #Music Taylor Swift Picks Up Five Teen Choice Awards - Great American...
Getting my galaxy note right now! ^_^	http://t.co/vUVLotId #AutoFollowBack Goo
Welcome to manchester united RVPglory glory Man.united~	Michael Jackson got at least 500 Awards, And 26 American Music Awards....#MJFact
I really like Owl City #amazing	Ex-Google exec's new venture helps students avoid corporate life /via@globeandmail
RT @Ci_Turner: I must admit I love my followers and my fan page @CiCisFanUpdate_ they really hold me down much love #AmericasGreatest #A ...	http://t.co/8pfNawc5
	CBS News - Barack Obama: Obama: Olympics welcome amid election tug-of-war . More #Obama #news - http://t.co/WYKc6tL5

Figure 2: Examples of subjective tweets and objective tweets

We discover several interesting characteristics of subjective authors. First, they tend to use pronouns frequently to represent themselves or others. Second, the opinionated lexicon words are used well for expressing their opinions. Third, they often use retweet to someone else's tweet expressing their agreement. On the contrary, the objective authors prefer to write longer informative messages and usually use links to provide more information. In fact, they are experts within a given domain, often belong to news agency, and tend to produce or propagate informative facts. Therefore, we found that the word usage of authors, the usage of links, and the length of messages are good features for estimating the subjectivity of tweet.

Secondly, we observe that we can classify subjective tweets and objective tweets among all hashtag related tweets. The following tweets show how hashtags are used in tweets.

- Big Bang Theory! *#love*
- *#18ThingsIWant* the iphone4s :)
- The new army in red devils is good! *#GGMU*
- *#News* Sylvia Woods' wake, funeral to celebrate life of restaurateur from Mount Vernon: A wa... *http://t.co/PhVjrMPm #MountVernon #NY #US*
- *RT @justinbieber*: What did u guys think?? I do this for y'all. #mybeliebers i love u. Thank u!!!

The hashtag '*#love*' in the first tweet is used instead of the word 'love' to emphasize the author's favour. The hashtag '*#18ThingsIWant*' used in the second tweet represents itself a positive opinion toward the 'iphone4s'. The hashtag '#GGMU', representing 'Glory Glory Manchester United', is used in the third tweet to express the author's favour toward the football team 'Manchester United'. The fourth tweet has several hashtags such as '*#News*'. '*#MountVernon*', '*#NY*', and '*#US*', and the hashtag '*#News*' may indicate the tweet is objective. Furthermore, it has the link "http://t.co.PhVjrMPm" which can be used to transmit the information. The fifth example expressed their opinions by using a retweet including a hashtag.

As shown in the example tweets, we discover that the word usage in hashtags and the subjectivity of the hashtag related tweets are useful for estimating the subjectivity of the original tweet.

Finally, we observe that the arrangement of words and the syntactic structure of reply tweets are very different between subjective tweets and objective tweets. Table 1 shows some original tweets and their replies.

Subjective tweets
I still *miss* Steve Jobs. :(@obviouslyben me *too*, man. @obviouslyben have *you* read his biography? @obviouslyben *i agree*.
@ElinGjones @htc are the best get @HTC one s its brilliant my htc won't *even* turn on, it's *shit*. *never* having one again @htc.
Welcome RvP to Manchester United :) *RT @Valencia7_ID* Welcome RvP to Manchester United :)
Objective tweets
USA holds over 10,000 nuclear bombs, Israel holds over 300 illegal nuclear bombs & the EU holds over 8,000, Iran holds 0. Is Iran a threat?! @persianfarzad does this tell more of the story too? it scores both nations' foreign interferences in the last century *[link]*

Table 1: Original tweets and their replies

We discover several characteristics in the replies of subjective tweets. First, there are more pronouns representing themselves or others in subjective tweets. We can find 'you', and 'i' in replies of first tweet. Second, subjective words are frequently used to express their opinions. The subjective word 'welcome' is used in the third tweet's reply to show a positive opinion toward 'RvP' (i.e. Robin Van Persie who is a soccer player). Third, retweets are often executed from subjective tweets. For example, the user 'Valencia7_ID' performed a retweet to welcome 'RvP' for joining 'Manchester United'. Fourth, many subjective words are used in subjective replies. In the first tweet's replies, the words 'too' and 'agree' are used to express sympathy with the original tweet. In the second tweet's replies, the words like 'even', and 'never' are used to emphasize the reply writer's opinion about the tweet, and slangs such as 'shit' is used to express his/her negative feeling toward the original tweet. On the other hand, links are used frequently in the objective tweets' replies in order to provide more information. As explained above, the word usage in replies, the usage of links, and the usage of retweets are good features for estimating the subjectivity of tweets.

We separate 30% of the manually labeled data for statistical analysis of our observational data. Here, we regard authors of subjective and objective tweets as subjective authors and objective authors, respectively. Similarly, we classify hashtags into subjective and objective hashtags, and replies into subjective replies (i.e. the replies to a subjective tweet) and objective replies (i.e. replies to an objective tweet). For statistical reliability, we remove authors who write both subjective and objective tweets and hashtags used in both subjective and objective tweets from the observational data.

Author-related Features	**Subjective Author**	**Objective Author**
Pronoun (words/collection)	0.071515	0.041261
Opinion (words/collection)	0.138744	0.115822
Link (links/tweet)	0.191576	0.540878
Retweet (retweets/tweet)	0.150191	0.086934
Tweet length (words/tweet)	13.0940	14.5949
Hashtag-related Features	**Subjective Hashtag**	**Objective Hashtag**
Pronoun (words/collection)	0.049459	0.038440
Opinion (words/collection)	0.109260	0.097661
Link (links/tweet)	0.4081324	0.60730968
Retweet (retweets/tweet)	0.15206397	0.1442972
Tweet length (words/tweet)	14.0582	14.73265
Reply-related Features	**Subjective Reply**	**Objective Reply**
Pronoun (words/collection)	0.077921	0.061157
Opinion (words/collection)	0.182479	0.151225
Retweet (retweets/tweet)	0.028091	0.018198
Link (links/tweet)	0.075089	0.170917
Discourse marker (words/collection)	0.073730	0.044302

Table 2: Statistical information for observational data including 15,973 tweets from subjective authors, 88,029 tweets from 729 objective authors, 132,502 tweets with 43 subjective hashtags, 684,869 tweets with 102 objective hashtags, 177 replies to 51 subjective tweets, and 467 replies to 104 objective tweets.

Table 2 presents statistical information for extrinsic characteristics of the observational data. In this table, we can observe that subjective authors tend to use more pronouns, opinionated words, and retweets than objective authors. On the contrary, objective authors tend to use more links and longer tweets than the subjective authors. With regard to hashtags, we can find that there are more pronouns, opinionated words, and retweets in subjective hashtags than in objective hashtags; however, there are more links and longer tweets in objective hashtags than in subjective hashtags. From the small amount of reply data, we observe that subjective replies tend to have more pronouns, opinionated words, retweets, and discourse markers than objective replies, whereas objective replies tend to have more hyperlinks than subjective replies.

4 Features for Opinion Retrieval Targeting Twitter

In this paper, we propose an opinion retrieval system that retrieves tweets according to topical relevance and subjectivity. The proposed system was developed based on a learning-to-rank framework, and many useful intrinsic and extrinsic features are identified for estimating topical relevance and subjectivity, and for ranking tweets. While intrinsic features are extracted from the original tweet, extrinsic features are

extracted from the expanded tweets related to the original tweet with respect to author, hashtag, or reply.

As we described in the last section, the author related tweets, hashtag related tweets, and the reply related tweets are especially good sources for estimating subjectivity. Furthermore, the expanded tweets are very helpful for alleviating data sparseness.

4.1 Intrinsic Features of Tweets

Opinion retrieval targeting tweets involves retrieving topically relevant tweets having subjective words. In this study, we first estimate topical relevance and subjectivity by using the intrinsic features of the original tweets. This estimation is the first baseline used in the opinion retrieval experiment.

- *Topical relevance* denotes how well a retrieved document or a set of documents topically meets the information needs of the user. The features for estimating topical relevance are related directly to the ranking model. In this study, we choose the language model represented in the Equation (1) for estimating topical relevance.

$$\log P(Q|D) = \sum_{i=1}^{n} \log \frac{f_{q_i,D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu} \quad (1)$$

  In this formula, Q denotes a query consisting of n query words, D denotes a document, $|D|$ is the length of the document, C denotes a collection of documents, $|C|$ is the total number of word occurrences in the collection, $f_{q_i,D}$ is the number of times a query word q_i occurs in the document D, c_{q_i} is the number of times a query word occurs in the collection of documents, and μ is a parameter whose value is set empirically.

- *Document subjectivity* denotes how subjectively a document is written. Previous studies estimated subjectivity based on the occurrence rate of subjective words in the documents ([Eguchi and Lavrenko 2006], [Gerani et al. 2011]). [Luo et al. 2012] extracted pseudo-subjective words and pseudo-objective words from a tweet corpus and used them for estimating the subjectivity of tweets. In this study, we used a publicly available subjective lexicon**.** and simply counted how many words of the original tweet are in the list as represented in Equation (2):

$$Score_D = \frac{\sum_{j=1}^{|L \cap D|} f_{w_j, L \cap D}}{|D|} \quad (2)$$

  In this formula, D denotes a set of words in the document D, L denotes the list of subjective words, $L \cap D$ denotes a set of common words in the document D and a subjective word set L, $|D|$ is the length of a document,

$|L \cap D|$ is the number of words in the set $L \cap D$ and $f_{w_j,L\cap D}$ is the number of times a word w_j occurs in the set $L \cap D$.

4.2 Extrinsic Features of Tweets

As stated in the section on new observations, we claim that author-related tweets, hashtag-related tweets, and the reply list of a tweet are good sources for estimating subjectivity. In this section, we introduce useful extrinsic author-related features, hashtag-related features, and reply-related features for estimating the subjectivity of tweets.

4.2.1 Features Derived from Author-Related Tweets

We identify useful features for classifying authors of tweets into subjective authors and objective authors. These include the occurrence rate of subjective words, using rate of pronouns, the occurrence rate of retweet, using rate of links, and the average length of tweets. These features are described in detail as follows:

- The *occurrence rate of subjective words* indicates how often subjective words are written by authors. Clearly, subjective authors tend to write more subjective words than objective author. In this study, we estimate the rate of subjective word use by employing Equation (3) as follows:

$$Score_{opword,Author} = \frac{\sum_{j=1}^{|L|} f_{l_j,A}}{|A|} \ (l_j \in L) \tag{3}$$

In this formula, L denotes the list of subjective words, A denotes a collection of tweets written by an author, $|A|$ is the total number of word occurrences in A, and $f_{l_j,A}$ is the number of times the word l_j, which is a member of L, occurs in A.

- The *occurrence rate of pronouns* indicates how often pronouns are written by authors. Subjective authors tend to write more pronouns than objective authors. There are several studies indicating that first-person pronouns (e.g. I, my, me, mine, we, our, etc.) are effective clues *for identifying* subjectivity [Atkinson et al. 2015]. In this study, we assume that not only first-person pronouns but also second-person pronouns (e.g. you, your, etc.) and third-person pronouns (e.g. he, she, it, etc.) can be helpful for identifying subjectivity of the tweet. Therefore, we include second-person pronouns and third-person pronouns when counting the number of pronouns. In this study, we estimate the rate of using pronouns by employing Equation (4) as follows:

$$Score_{pronoun,Author} = \frac{\sum_{j=1}^{|P|} f_{p_j,A}}{|A|} \ (p_j \in P) \tag{4}$$

In this formula, P denotes a pronoun list, $|P|$ is the number of pronouns, $f_{p_{j,A}}$ denotes the frequency of the pronoun p_j, which is a member of P, occurred in A.

- The *rate of retweets* indicates how often an author retweets others' tweets. A retweet in a format such as "RT @[ID]: xxxx" is the way that the author agrees with the original tweet. Subjective authors tend to use more retweet function than objective authors. In this study, we estimate the rate of retweets by employing Equation (5) as follows:

$$Score_{retweet,Author} = \frac{\sum_{j=1}^{N_A} retweet_{a_j}}{N_A} \quad (5)$$

In this formula, N_A denotes the number of tweets that the author writes, a_j denotes a tweet in A, and $and \sum_{j=1}^{N_A} retweet_{a_j}$ is the number of retweets among the tweets written by the author A.

- The *occurrence rate of links* indicates how often links are written by authors. The link feature was used in previous studies of opinion mining which deals with individual tweet ([Go et al. 2009], [Barbosa and Feng 2010], [Luo et al. 2012]). In Twitter, news accounts and experts in their fields have left links to fill in missing information because of short document length. Therefore, objective authors tend to use more link in their tweets than subjective authors. In this study, we estimate the rate of using links by employing Equation (6) as follows:

$$Score_{link,Author} = \frac{\sum_{j=1}^{N_A} link_{a_j}}{N_A} \quad (6)$$

In this formula, $link_{a_j}$ is the number of links in the tweet a_j.

- *Tweet average length* indicates how long tweets are for a given author. Previous studies showed that the length of tweets is an effective feature because of information gain[Luo et al. 2012]. People who usually write informative tweets tend to write longer tweet to provide information. In contrast, people writing their opinion statements tend to write their thoughts briefly but do not provide any further information. In this study, we estimate tweet average length by employing Equation (7) as follows:

$$Score_{AvgLen,Author} = \frac{\sum_{j=1}^{N_A} doclen_{a_j}}{N_A} \quad (7)$$

In this formula, $doclen_{a_j}$ is the document length of the tweet a_j.

4.2.2 Features Derived from Hashtag-Related Tweets

In this section, we introduce useful extrinsic hashtag-related features. These include using the occurrence rate of subjective words, using the occurrence rate of pronouns, the occurrence rate of retweet, using rate of links, and the average length of tweets. These features are described in detail as follows:

- The *occurrence rate of subjective words* indicates how often subjective words are used in hashtag-related tweets and in hashtags themselves. Sometimes, subjective words are included in the words in a hashtag such as "*Big Bang Theory! #love"*, or the hashtag partially contains a subjective word such as in "*#18ThingsIWant the iphone4s :)*" In this study, we calculate the rate of using subjective words by employing Equation (8) as follows:

$$Score_{opword,Hashtag} = \frac{\sum_{j=1}^{|L \cup OH|} f_{l_j,H}}{|H|} \ (l_j \in L \cup OH) \quad (8)$$

  In this formula, In this formula, OH denotes the list of subjective hashtags containing subjective words within them such as '#18thingsILike' or '#love', $L \cup OH$ denotes the union of the subjective hashtag list and the subjective word list, H denotes a collection of tweets where a hashtag is used, $|H|$ is the number of words in H, and $f_{l_j,H}$ is the number of times the word l_j, which is a member of $L \cup OH$, occurs in H.

- The *occurrence rate of Pronouns* indicates how often pronouns are used in hashtag-related tweets. Pronouns are sometimes used to express the authors themselves in subjective hashtags such as in "*I #love galaxy note!!*", or "*I love my kindle fire >>>> #loveit*". As implied in the examples, the using rate of pronouns can be a good feature from which to estimate the subjectivity of each hashtag. In this study, we calculate the rate of using pronouns by employing Equation (9) as follows:

$$Score_{pronoun,Hashtag} = \frac{\sum_{j=1}^{|P|} f_{p_j,H}}{|H|} \ (p_j \in P) \quad (9)$$

  In this formula, $f_{p_j,H}$ is the number of times the pronoun p_j, which is a member of P, occurs in H.

- The *rate of retweets* indicates how often retweets are used where the hashtags are used in tweets. Retweets are used more often when the tweets have subjective hashtags than when they have objective hashtags. In this study, we estimate the rate of using retweets by employing Equation (10) as follows:

$$Score_{retweet,Hashstag} = \frac{\sum_{j=1}^{N_H} retweet_{h_j}}{N_H} \quad (10)$$

In this formula, N_H denotes the number of tweets where the hashtag is used, and $\sum_{j=1}^{N_H} retweet_{h_j}$ denotes the number of retweets against the tweets with a hashtag.

- The *occurrence rate of links* indicates how often links are used in hashtag-related tweets. Links are used more often when the tweets have objective hashtags than when they have subjective hashtags. In this study, we estimate the rate of using links by employing Equation (11) as follows:

$$Score_{link,Hashtag} = \frac{\sum_{j=1}^{N_H} link_{h_j}}{N_H} \quad (11)$$

In this formula, $link_{h_j}$ denotes the number of links used in the tweet with hashtag.

- *Tweet average length* indicates the length of a tweet with a hashtag. The average length of informative tweets with hashtags is longer than that of subjective tweets with hashtags. In this study, we estimate the tweet average length by employing Equation (12) as follows:

$$Score_{AvgLen,Hashtag} = \frac{\sum_{j=1}^{N_H} doclen_{h_j}}{N_H} \quad (12)$$

In this formula, $doclen_{h_j}$ is the document length of the tweet with a hashtag.

4.2.3 Features Derived from Replies to a Tweet

We introduce useful features for estimating the subjectivity of replies in the set of reply tweets. They include using the occurrence rate of subjective words, the occurrence rate of pronouns, the rate of retweets, the occurrence rate of links, and the occurrence rate of subjective discourse markers. These are described in detail as follows:

- The *occurrence rate of subjective word* indicates how often subjective words are used when people write replies about a tweet. Replies to a subjective tweet typically contain more subjective words than replies to an objective tweet. In this study, we estimate the rate of using subjective words by employing Equation (13) as follows:

$$Score_{opword,Reply} = \frac{\sum_{j=1}^{|L|} f_{l_j,R}}{|R|} \ (l_j \in L) \quad (13)$$

In this formula, R denotes a collection of replies to the original tweet, $|R|$ is the number of words in R, and $f_{l_{j,R}}$ is the number of times the word l_j, which is a member of L, occurs in R.

- The *occurrence rate of Pronoun* indicates how often pronouns are used when people write replies about a tweet. Replies to a subjective tweet typically

contains more pronouns than replies to an objective tweet. In this study, we estimate the rate of using pronouns by employing Equation (14) as follows:

$$Score_{pronoun,Reply} = \frac{\sum_{j=1}^{|P|} f_{p_{j,R}}}{|R|} \ (p_j \in P) \quad (14)$$

In this formula, $f_{p_{j,R}}$ is the number of times the pronoun p_j, which is a member of P, occurs in R.

- The *rate of retweets* indicates how often retweets are used when people write replies about a tweet. The replies to a subjective tweet have more retweets than replies to an objective tweet. In this study, we estimate the rate of using retweet by employing Equation (15) as follows:

$$Score_{retweet,Reply} = \frac{\sum_{j=1}^{N_r} retweet_{r_j}}{N_r} \quad (15)$$

In this formula, N_r denotes the number of replies in R, r_j denotes a tweet in R, and $\sum_{j=1}^{N_r} retweet_{r_j}$ denotes the number of retweets of the whole replies.

- The *occurrence rate of links* indicates how often links are used when people write replies about a tweet. In Twitter, the main role of links is to fill in the missing information. Replies to an objective tweet have more retweets than replies to a subjective tweet. In this study, we estimate the rate of using links by employing Equation (16) as follows:

$$Score_{link,Reply} = \frac{\sum_{j=1}^{N_r} link_{r_j}}{N_r} \quad (16)$$

In this formula, $link_{r_j}$ is the number of link occurring in tweet r_j.

- The *occurrence rate of subjective discourse markers* indicates how often subjective discourse markers are used when people write replies about a tweet. We define the subjective discourse markers as words used in subjective statements. Replies of subjective tweet have more subjective discourse markers than replies to an objective tweet. In this study, we estimate the rate of using subjective discourse markers by employing Equation (17) as follows:

$$Score_{discourse,Reply} = \frac{\sum_{j=1}^{|DW|} f_{d_{j,R}}}{|R|} \ (d_j \in DW) \quad (17)$$

In this formula, DW denotes a list of subjective discourse markers, $|DW|$ is the number of words in the list DW, and $f_{d_{j,R}}$ is the number of times the discourse marker d_j, which is a member of DW, occurs in R.

The use of discourse markers in reply tweets can be a useful clue in recognizing the writer's intent, and thus in estimating subjectivity.

Writer's Intent	**Discourse Markers**
Argumentation	but, because, why
Emphasis	never, ever, forever
Contradiction	dnt, dunno
Degradation	as*x*, f*xx*king, n*xx*ga, lol, lmao, s*x*it
Onomatopoeia	yay, hmmm, oooh

Table 3: Examples of using discourse markers

Discourse markers are used for many purposes such as argumentation, emphasis, contradiction, etc. representing the writer's intent. Examples of discourse markers are shown in Table 3. In this study, we automatically extract discourse markers used in subjective statements by employing the chi-square method [Luo et al. 2012], and we extract them from the observation data as stated in Section 3. As a result, we collect 136 discourse markers that are useful features for estimating the subjectivity of reply tweets.

5 Experiments

5.1 Experimental Setup

5.1.1 Evaluation Corpus

We constructed a Twitter Corpus by crawling 280,208,546 tweets written in English during one month from July 24th of 2012 to August 23rd of 2012. The constructed corpus became the target of the opinion retrieval system. In this experiment, we employed a technique called pooling to collect the top 100 results of each query from the rankings obtained by three different retrieve model: the vector space model[2], BM25[3], and the language model[4].

Relevance judgments are performed by human experts from Amazon Mechanical Turk which is a popular crowdsourcing services. They select one of three choices:

(1) The tweet is topically relevant and subjective
(2) The tweet is topically relevant but objective

[2] The vector space model is an algebraic model for representing documents and queries as vectors of terms Documents can be ranked by computing the distance between the points representing the documents and the query [Salton et al. 1975].
[3] BM25 is a ranking function used by a search engine to rank matching documents according to their relevance to a given search query based on a probabilistic framework [Robertson et al. 1994].
[4] A language model is a probability distribution over sequences of words. Documents can be ranked based on the probability of the query in the document language model [Croft et al. 2010].

(3) The tweet is topically irrelevant.

The category distributions are as follows: (1) Relevant and subjective: 870 tweets (17.4 tweets per query), (2) Relevant and objective: 1,172 tweets (23.44 tweets per query), (3) Irrelevant: 1,629 tweets (32.58 tweets per query). Tweets designated in category (1) are appropriate, desired documents for the opinion retrieval task. Five American annotators select one of three choices for the top 100 tweets of each query. We regard the tweet as the correct answer when three or more annotators agreed to select the same choice.

In this study, we used the same queries as in [Luo et al. 2012] for comparison purposes. The total number of queries was 50, and the average length of the queries was about 1.94 words. The 50 queries covered various topics such as organizations (e.g., Manchester United), products (e.g., MacBook Pro), people (e.g., Bill Gates), locations (e.g., Iran), movies (e.g., Big Bang), and others (e.g. speech recognition).

5.1.2 Evaluation Methodology

In this experiment, we used mean average precision (MAP)[5], precision@5 (P@5)[6] and precision@10 (P@10) to evaluate the effectiveness of the proposed opinion retrieval system. MAP is used to evaluate the overall performance of the proposed system, and P@5 and P@10 are used to evaluate the performance of the highly ranked results list.

5.2 Experimental Results

In order to compare the effectiveness of the proposed system using extrinsic features, two different baselines employing different sets of features were used in this experiment.

1) The basic system using only primitive features: a language model score and the rate of using subjective words are the only features used for estimating the topical relevance and subjectivity of tweet messages as in the previous opinion retrieval systems targeting blogs ([Gerani et al. 2011], [Lee et al. 2012]).
2) [Luo et al. 2012]: This state-of-the-art opinion rerieval system targeting Twitter is based on [Luo et al. 2012]. This system tries to utilize intrinsic features of original tweets such as hashtags and hyperlinks and meta information, such as the number of followers and the number of tweets, in addition to query-relevance features and word-based subjectivity features introduced in previous studies.

The RankSVM of the SVM Light [Joachim 2002] toolkit was applied for the proposed opinion retrieval system, and 10-fold cross validation was used to avoid overfitting of the training data. When we performed 10-fold cross validation, we

[5] Mean average Precision (MAP) is the mean of the average precision scores for each query [Croft et al. 2010].
[6] Precision@5 refers the precision score at the fifth rank [Croft et al. 2010].

divided 50 queries into 10 folds so that each fold contained 5 queries. We used MPQA Lexicon [Wilson et al. 2005] for the subjective word list. In this study, we developed a mechanism of saving feature information in a hash table so that the proposed system can directly use the feature information in real time. Consequently, the proposed opinion retrieval system runs in real time like a general information retrieval system, and has an advantage in terms of efficiency.

5.2.1 Effectiveness of Extrinsic Features

In this experiment, the performance of the opinion retrieval system was compared with the performance of the two baselines with respect to MAP, P@5, and P@10. The six opinion retrieval systems listed in Table 4 are different according to the features used in each system, as shown in Table 5. The basic system uses only intrinsic features such as language model score, and document subjectivity. As shown in Table 5, [Luo et al. 2012] additionally uses Twitter-specific features and author meta features such as the presence of hashtags, and the number of followers. The features of the proposed system are extended to include author-related features, hashtag-related features, and reply-related features.

	MAP	**P@5**	**P@10**
Basic System	0.3122	0.3378	0.3311
[Luo et al. 2012]	0.3631	0.4133	0.3978
Proposed (All)	**0.3898▲**	**0.4356▲**	**0.4067▲**
Proposed -Author	0.3337△	0.3778▲	0.3600△
Proposed -Hashtag	0.3705▲	0.4000▲	0.3978▲
Proposed - Reply	0.3733▲	0.4178▲	0.4022▲

Table 4: Performance comparison between the proposed and baseline systems. Black triangles indicate the cases where the p-value is less than 0.01, and white triangles indicate the case when the p-value is less than 0.05.

As shown in Table 4, the proposed system significantly outperformed the two baselines. The performance of the proposed system is much better than the first baseline, and thus we can say that this performance improvement proves the effectiveness of the proposed method. In particular, we determine that the author feature is most effective, but that the hashtag and the reply features are also effective. Interestingly, our proposed system outperformed the state-of-the-art system [Luo et al. 2012] even though we did not include latter's Twitter-specific features and author meta features, as shown in Table 5. This result implies that the proposed extrinsic features are more useful than the Twitter-specific features used by [Luo et al. 2012].

Feature		Basic System	[Luo et al. 2012]	Proposed
Topical relevance		○	○	○
Document Subjectivity		○	○	○
[Luo et al. 2012]	Hashtag Y/N	×	○	×
	Tweet Number	×	○	×
	Follower Number	×	○	×
	List Number	×	○	×
	Friend Number	×	○	×
Author	Rate of Subjective Word	×	×	○
	Rate of Pronoun	×	×	○
	Rate of Link	×	×	○
	Rate of Retweet	×	×	○
	Tweet Length	×	×	○
Hashtag	Rate of Subjective Word	×	×	○
	Rate of Pronoun	×	×	○
	Rate of Link	×	×	○
	Rate of Retweet	×	×	○
	Tweet Length	×	×	○
Reply	Rate of Subjective Word	×	×	○
	Rate of Pronoun	×	×	○
	Rate of Link	×	×	○
	Rate of Retweet	×	×	○
	Rate of Subjective Discourse Marker (DCMarker)	×	×	○

Table 5: List of features used in the baselines and the proposed system

We also performed an ablation experiment in order to show that author-related features, hashtag-related features, and reply-related features are useful for improving performance. A paired t-test was used for statistical validation The results of the ablation experiments also show that author related features can achieve the greatest performance improvement. In addition, the performance results of *Proposed – Hashtag* and *Proposed – Reply*, show that the hashtag-related features and the reply-related features are also useful for improving the performance of the proposed opinion retrieval system.

We also performed the ablation experiment in order to evaluate the effects of each feature for opinion retrieval. The experimental results of using 15 different features are shown in Table 6.

Among author related features, pronouns appear to be the best because the results of the Proposed-Pronoun have the lowest value with respects to MAP. We assume that the pronoun feature is important because it is used to express the opinions of authors. As shown in Table 6, the performance of Proposed-Link and Proposed-Retweet are relatively poor with respect to MAP. We determine that the link feature shows that a tweet is informative and objective, and the retweet feature is also important because it is used as a tool for expressing the thought of the author. Additionally, the length of the tweet and subjective words are good features for

estimating subjectivity. Furthermore, we can conclude that pronoun use and retweet frequency are useful features for opinion retrieval compared to other features.

	Features	**MAP**	**P@5**	**P@10**
	Proposed	0.3898	0.4356	0.4067
Author Feature	Proposed – Subjective word	0.3793	0.4178	0.4089
	Proposed – Pronoun	**0.3716**	0.4267	0.3978
	Proposed – Retweet	0.3788	0.4311	0.4022
	Proposed - Tweet Length	0.3796	0.4444	0.4111
	Proposed - Link	0.3788	0.4000	0.3956
Hashtag Feature	Proposed - Subjective word	**0.3778**	0.4222	0.4111
	Proposed - Pronoun	0.3796	0.4133	0.3956
	Proposed - Retweet	0.3803	0.4311	0.4022
	Proposed – Tweet Length	0.3790	0.4267	0.4089
	Proposed - Link	0.3803	0.4311	0.4022
Reply Feature	Proposed - Subjective word	0.3809	0.4178	0.4111
	Proposed - Pronoun	0.3840	0.4311	0.4089
	Proposed - Link	0.3824	0.4222	0.4111
	Proposed - Retweet	0.3813	0.4267	0.4022
	Proposed – DCMarker	**0.3774**	0.4133	0.4022

Table 6: Performance of ablation experiment of Proposed

When we compare the results of *Proposed – Author* in Table 4 with each result of excluding individual author-related features presented in Table 6, the performance of Proposed – Author has the lowest value with respect to MAP, P@5, and P@10. This fact may imply that the combination of features can create a synergy effect in opinion retrieval.

Furthermore, we find that each hashtag-related feature can improve the performance of the opinion retrieval system. We conclude in particular that the use of subjective words within the hashtag (e.g. '#love', '#18ThingsIWant', etc.) can be a strong clue for estimating the subjectivity of related tweets. In addition, pronouns, retweets, tweet length, and link features are helpful for estimating subjectivity.

Among reply-related features, discourse markers turns out to be the best because the results of *Proposed – DCMarker* have the lowest value with respect to MAP. The performance result of each reply-related feature shows that identifying specific words such as subjective words, pronouns, and discourse markers is an important task in defining useful features for opinion retrieval.

5.2.2 Improvement of State-of-the-Art Performance

In order to verify that our proposed author-related features, hashtag-related features, and reply-related features have a positive effect on the state-of-the-art opinion retrieval system targeting Twitter, we try to add all of our proposed features into the [Luo et al. 2012]'s system. The experimental results are presented in Table 7.

	MAP	**P@5**	**P@10**
[Luo et al. 2012]	0.3631	0.4133	0.3978
Luo + Proposed	**0.3968▲**	**0.4533△**	**0.4089△**
Luo + Proposed – Author	0.3743	0.4356△	0.4089△
Luo + Proposed – Hashtag	0.3838	0.4356△	0.4022
Luo + Proposed - Reply	0.3794△	0.4222	0.4067

Table 7: Performance of including our proposed features into [Luo et al. 2012]'s framework. Black triangles indicate the cases where the p-value is less than 0.01, and white triangles indicate the case when the p-value is less than 0.05.

Table 7 shows that *Luo + Proposed* (i.e.. the system using all of Luo's features and all of our proposed extrinsic features) is much better than the performance of [Luo et al. 2012] which is the state-of-the-art opinion retrieval system targeting Twitter with respect to MAP, P@5, and P@10. The statistical significance test is also performed by using a paired t-test.

This experiment shows that the performance of the state-of-the-art opinion retrieval system targeting Twitter can be improved significantly by using the proposed extrinsic features, and that the features in [Luo et al. 2012] are not sufficient, especially in estimating the subjectivity of tweets.

When we remove author-related features from the *Luo + Proposed* system, performance degradation is largest with respect to MAP. This fact indicates that author-related features have the strongest positive effect on the performance improvement of the state-of-the-art system. Table 7 also shows that hashtag-related features and reply-related features have positive effects on performance improvement.

6 Conclusion

In this paper, we investigated the effects of including extrinsic features of tweets on the performance of an opinion retrieval system. We proposed author-related extrinsic features, the hashtag-related extrinsic features, and the reply-related extrinsic features. Experimental results show that all of the proposed features are useful in the opinion retrieval system, that they improve the performance of the state-of-the-art opinion retrieval system, and that they can alleviate the data sparseness problem which is the major obstacle in estimating the subjectivity of tweets.

Opinion retrieval targeting Twitter aims to retrieve subjectively written, relevant tweet documents in real time. If a good opinion retrieval system targeting Twitter is developed, it can be utilized for discerning customer opinions in real time and for quickly establishing a marketing strategy reflecting customer opinions.

7 Future Work

In future work, we will try to find useful extrinsic tweet sources for estimating topical relevance and to develop an opinion retrieval model which can consider both original tweet sources and extrinsic tweet sources.

Acknowledgements

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Plannig (2012M3C4A7033344).

References

[Atkinson et al. 2015] Atkison,J., Sals,G., Figueroa,A.: Improving Opinion Retrieval in Social Media by Combining Feature-based Coreferencing and Memory-based Learning, *Information Science*, 20-31, 2015

[Barbosa and Feng 2010] Barbosa,L., Feng, J.: Robust Sentiment Detection on Twitter from biased and noisy data, *Proceedings of the 23rd International Conference on Computational Linguistics*, 36-44, 2010

[Croft et al. 2010] Croft,W.B., Metzler,D., Strohman,T.: *Search Engines: Information Retrieval in Practice. Pearson Education Inc.*, 2010

[Eguchi and Lavrenko 2006] Eguchi,K., Lavrenko,V.: Sentiment Retrieval using Generative Model, *Proceedings of the 2006 conference on Empirical Methods in Natural Language Processing*, 345-354, 2006

[El-Ganiany et al. 2013] El-Ganiany,T., Wei,Z., Magdy,W., Gao,W.: Overview of the TREC-2013 Microblog Track, *Proceedings of the 21th Text REtrieval Conference*, USA, 2013

[El-Ganiany et al. 2014] El-Ganiany, T., Magdy,W., Rafea,A.: Hyperlink-Extended Pseudo Relevance Feedback for Improved Microblog Retrieval., *Proceedings of the first international workshop on Social media retrieval and analysis*, 7-12, 2014

[Gerani et al. 2009] Gerani S., Carmen, M.J., Crestani,F.: Investigating Learning Approaches for Blog Post Opinion Retrieval, *31th European Conference on IR Research*, 313-324, 2009

[Gerani et al. 2011] Gerani,S., Carmen, M.J., Crestani,F.: Proximity-based Opinion Retrieval, *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 403-410, 2010

[Go et al. 2009] Go, A., Bhayani,R., Huang,L.: Twitter Sentiment Classification using Distant Supervision, *Technical Report, Stanford Digital Library Technologies Project*, 2009

[Huang and Croft 2009] Huang,X., Croft,W., A Unified Relevance Model for Opinion Retrieval, *Conference on Information and Knowledgement Management (CIKM)*, 947-956, 2009

[Joachim 2002] Joachim,T.: Generating Typed Dependency Parses from Phrase Structure Parses, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*,133-142, 2002

[Lee et al. 2012] Lee,S.W., Song,Y.I., Han,K.S., Rim,H.C.: A New Generative Opinion Retrieval Model Integrating Multiple Ranking Factor, *Journal of Intelligent Information System*, 487-505, 2012

[Lin et al. 2014] Lin,J., Efron,M., Wang,Y.,Sherman,G.: Overview of the TREC-2014 Microblog Track. *Proceedings of the 22th Text REtrieval Confernece*, USA, 2014

[Luo et al. 2012] Luo, Z., Osborne, M., Wang, T.: Opinion Retrieval in Twitter, *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 507-510, 2012

[Luo et al. 2013] Luo, Z., Tang, J, Wang,T.: ,Propagated Opinion Retrieval in Twitter, 16-28, *WISE*, 2013

[Macdonald et al. 2007] Macdonald,C., Ounis,I., Soboroff,I.: Overview of the TREC-2007 Blog Track, *Proceedings of the Fifteenth Text REtrieval Conference*, USA, 2007

[McCredie and Macdonald 2013] McCreadie,R., Macdonald,C.: Relevance in Microblogs: Enhancing Tweet Retrieval using Hyperlinked Documents., *Proeedings of the 10th Conference on Open Research Areas in Information Retrieval*, 189-196, 2013

[Onal et al. 2015] Onal,K.D., Altingovde,I.S., Karagoz,P.: Utilizing Word Embeddings for Result Diversification in Tweet Search, *11th Asia Information Retrieval Societies Conference, AIRS 2015*, Brisbane, QLD, Australia, December 2-4, 2015. Proceedings, 366-378, 2015

[Ounis et al. 2006] Ounis,I., Rijke, M., Macdonald,C., Mishne,G.:Overview of the TREC 2006 Blog Track, *Proceedings of the Fifteenth Text REtrieval Conference*, USA, 2006

[Ounis et al. 2008] Ounis,I., Macdonald,C., Soboroff,I.: Overview of the TREC-2008 Blog Track, *Proceedings of the Fifteenth Text REtrieval Conference*, USA, 2008

[Ounis et al. 2011] Ounis,I., Macdonald,C., Soboroff,I.: Overview of the TREC-2011 Microblog Track. *Proceedings of the Nineteenth Text REtrieval Conference*, USA, 2011

[Rao et al. 2015] Rao,J., Lin,J., Efron,M.: Reproducible Experiments on Lexical and Temporal Feedback for Twitter Search, *37th European Conference on IR Research, ECIR 2015*, Vienna, Austria, March 29 - April 2, 2015. Proceedings, 755-767, 2015

[Robertson et al. 1994] Robertson,S.E., Walker,S., Jones,S., Hancock-Beaulieu,M.M.: Okapi at TREC-3, *Proceedings of the Third Text Retrieval Conference*, 1994

[Salton et al. 1975] Salton,G., Wong,A., Yang,C.S..: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, vol. 18, nr. 11, 613-620, 1975

[Soboroff et al. 2012] Soboroff,I., Ounis,I., Macdonald,C., Lin,J.: Overview of the TREC-2012 Microblog Track, *Proceedings of the Twentieth Text REtrieval Conference*, USA, 2012

[Saif et al. 2014] Saif,H., He,Y., Fernandez,M., Alani,H.: Semantic Patterns for Sentiment Analysis of Twitter, *13th International Semantic Web Conference*, 324-340, 2014

[Severyn et al. 2015] Severyn,A., Moschitti,A.: Twitter Sentiment Analysis with Deep Convolutional Neural Network, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 373-382, 2015

[Tang et al. 2015] Tang,J., Nobata,C. Dong,A., Liu,H., Propagation-based Sentiment Analysis for Microblogging Data, *Proceedings of the 2015 SIAM International Conference on Data Mining*, 577-585, 2015

[Wilson et al. 2005] Wilson,T., Wiebe,J., Hoffmann,P., Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347-354, 2005

[Zhang and Yu 2006] Zhang, W., Yu, C., UIC at TREC 2006 Blog Track, In TREC 2006: *Proceedings of the fifteenth Text REtrieval Conference*, USA, 2006

[Zhang and Meng 2007] Zhang,W., Meng,W.: Opinion Retrieval from Blogs, *Conference on Information and Knowledgement Management (CIKM)*, 831-840, 2007

[Zhang and Ye 2008] Zhang,M., Ye,X., A generative Model to Unify Topical Relevance and Lexicon-based Sentiment for Opinion Retrieval, *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 411-418, 2008

[Zhang et al. 2013] Zhang, J., Minami,K., Kawai,Y., Kumamoto,T.: Personalized Web Search Using Emoticon Features, *Availibility, Reliability, and Security in Information Systems and HCI*, 69-83, 2013