Detecting Comments on News Articles in Microblogs

Alok Kothari, Walid Magdy, Kareem Darwish, Ahmed Mourad, and Ahmed Taei

Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar

{akothari, wmagdy, kdarwish, amourad, ataei}@qf.org.qa

Abstract

A reader of a news article would often be interested in the comments of other readers on anarticle, because comments give insight into popular opinions or feelings toward a given piece of news. In recent years, social media platforms, such as Twitter, have become a social hub for users to communicate and express their thoughts. This includes sharing news articles and commenting on them. In this paper, we propose an approach for identifying "comment-tweets" that comment on news articles. We discuss the nature of comment-tweets and compare them to subjective tweets. We utilize a machine learning-based classification approach for distinguishing between comment-tweets and others that only report the news. Our approach is evaluated on the TREC-2011 Microblog track data after applying additional annotations to tweets containing comments. Results show the effectiveness of our classification approach. Furthermore, we demonstrate the effectiveness of our approach on live news articles.

Introduction

The last decade has witnessed a major decline in print newspaper readership in favor of online newspapers and news sites. Aside from the convenience and freshness of online news sites, many users are drawn by the ability to comment and express their feelings and opinions on the news. Such commentary gives insights about the readership's opinion, thoughts, and sentiment and provides a forum for discussing the news.

Commentary on the news has spilled over into social media, with users either: sharing news articles (via URL or short URL) and commenting on them; or commenting about a topic in the news without a reference to a specific article. One popular social media platform for sharing and commenting on news is Twitter. The nature of comments on Twitter differs considerably from those on news sites. They differ in the following way:

1. *Nature and Length*: Since Twitter users are constrained by the short length of tweets, comments are typically more focused. The comment can be just a word, an abbreviation, an emoticon, a hashtag, or a question.
2. *Size and diversity*: the number of tweets commenting on a given news article can be much larger than the number of comments on the website reporting the news itself. Unlike commentary on news sites that reflects the opinion of the readership of the news source, commentary on Twitter is made by a much wider audience and is potentially more reflective of public opinion.
3. *Varying importance of comments*: Some tweets are more noteworthy than others, particularly if well-known persons author them. Such tweets are typically circulated via "re-tweeting". Highly re-tweeted tweets or those authored by persons with many followers would perhaps be more interesting to users.
4. *Freshness*: Often users start tweeting and commenting on news even before any news article is published.

In this paper, we propose the task of identifying comments on specific news articles from Twitter. Such commentary can be provided alongside online news articles for an improved reader experience. The proposed task would provide comments on news articles that are succinct, diverse, and fresh.

A Tweet containing a comment, or "comment-tweet", provides a user's response to a news item. A comment-tweet may express sentiment/opinion, a question, a rumor, or a call to action. Comment-tweets are not those tweets that just restate or rephrase the news. Example comment-tweets are provided in Table 1. A comment can be expressed using a full sentence describing a user reaction, a simple emoticon, or a hashtag (#tag). Comment-tweets are different from subjective tweets. For example, though the first tweet in Table 1 would probably get a neutral sentiment, it is a comment tweet. Also, the last tweet would probably get a negative sentiment, but it would not count as a comment tweet, because it is the first sentence in the article. Thus, it is not a comment.

Table1. Example of comment and non-comment-tweets for a news article

New article	Protests in US over slow power restoration (http://www.aljazeera.com/news/americas/2012/11/2012111124337353188.html)
Comment-tweet	Gary Owen @ElSnarkistani: First world, meet the third world: Protests in US over slow power restoration http://feedly.com/k/Saz3Ds
	Beehorla Oba! @byolar2u: "@AJEnglish: Protests in US over slow power restoration http://aje.me/TSYnln" can this happen in 9ja, who sees you...???
Non-comment-tweet	@Em_Fawzy: Frustrated residents express anger outside Long Island utility over electricity outages caused by the storm Sandy. http://aje.me/WS0rxE

Given a set of tweets that are relevant to a news item, we applied a machine learning technique to classify them as comment-tweets or not. Relevant tweets for an article can be obtained through: finding retweets of the article; searching for tweets that are topically related to the article; or using other methods such as filtering. According to our data, a minority of relevant tweets are in fact comment-tweets. We used the TREC-2011 Microblog track data for creating a new test set to evaluate our approach for the proposed task. We identified a relevant article for every topic, and we ascertained comment-tweets from among the relevant tweets for the article. Our experimental results show the effectiveness of our approach for identifying comment-tweets with high precision and acceptable recall. Moreover, we validated the effectiveness of our developed system on a set of tweets that link to news articles showing the effectiveness of comment-tweets identification.

The contributions of this paper are as follows:

1. We propose the novel task of automatically detecting users' comments on news items from Twitter, which can be presented alongside online news articles for a better reader experience.
2. We define comment-tweets and show how they differ from subjective tweets.
3. We applied a classification approach using a large set of features for comment-tweets identification.
4. We built a manually annotated gold-standard dataset for the task, which is based on the TREC-2011 Microblog track dataset[1].

The rest of the paper is organized as follows: Section 2 reviews related work; Section 3 defines comment-tweets and describes their nature; Section 4 describes our approach for comment identification; Section 5 explains experimental setup; Section 6 analyses the subjectivity of comment-tweets on our test set, and demonstrates the differences experimentally; Section 7 reports the results and highlights the success of the approach on a set of 15 live news articles; Finally, Section 8 concludes the paper and provides possible future directions.

[1]http://cluster.qcri.org/~wmagdy/resources.htm

Related Work

Microblog Text Classification

Much research work has focused on classifying tweets into several classes for different uses. Most work focused on applying subjectivity and sentiment analysis to determine if a tweet is subjective and whether it has positive or negative sentiment (Barbosa and Feng 2010; Chen et al., 2012; Kilner and Hoadley, 2005; Jiang et al., 2011; Lin et al., 2011; Zhao et al., 2011). Subjective tweets are those that present an opinion on something or someone, while objective tweets are those that carry facts such as news or information (Liu, 2009). Other work applied additional classification to tweets (González-Ibáñez et al., 2011; Sriram et al., 2010). In (Sriram et al., 2010), tweets were classified into four categories: news, events, opinion and deals. González-Ibáñez et al. (2011) presented an approach for automatically detecting sarcastic tweets. Different approaches were suggested for tweet classification. The most common approach uses supervised learning, with support vector machines (SVM) and Naïve Bayesian classification being common approaches (Barbosa and Feng, 2010; González-Ibáñez et al., 2011; Jiang et al., 2011; Sriram et al., 2010). Such approaches focus on extracting features from tweet text and metadata, such as author information. Other approaches use generative models based on Latent Dirichlet Allocation (LDA) to identify subjective tweets (Lin et al., 2011; Wang et al., 2012). They used terms and their co-occurrences within different classes as features. Recent work studied the retrieval of opinionated tweets against a given topic and modeled the task as a ranking problem (Luo et al., 2012), where a machine learning-based ranking algorithm was applied to rank tweets according to both relevance and subjectivity simultaneously.

Microblog and News

Many online news providers are currently moving towards providing more interaction between news readers. For example, many news websites enable users to comment on news articles and to share articles on social media. In many news websites, comments below news articles are linked directly to readers' accounts on the social media sites, such as Facebook or Twitter. With the increased interactivity on the news sites and social networks, it is important to manage this interactivity and to use it to improve the news readers' experience.

Previous work has found solid evidence of the journalistic value of comments, including adding perspectives, insights, andpersonal experiences that can enrich a news story (Diakopoulos and Naaman, 2011; Gilbert and Karahalios, 2010; Kilner and Hoadley, 2005). The work by (Kilner and Hoadley, 2005) studied the motivations of users commenting on news articles. The

study shows that users typically comment on online news to: ask and answer questions, add information, share personal experience, express sentiment, and/or criticize or support the news. This study provides important insight for us in defining what constitutes a comment-tweet, where it shows that comment-tweets may not convey sentiment. Rather, comment-tweets extend to other forms.

Additional research has focused on coupling news and tweets. Subasic and Berendt (2011) and Zhao et al. (2011) used tweets as a news source and compared tweets to online news media to detect features for automatic news detection in tweets. Phelan et al. (2011) used tweets to recommend news articles to users based on their preferences. Yang et al. (2011) applied summarization to webpages using a model that selects sentences based on social-context of tweets linking to a news article. Gao et al. (2012) applied cross-media summarization to news articles and tweets to identify complementary information from both on the same news.

In our work, we propose a novel task for coupling online news and tweets to enrich readers' experience when reading online news by providing users' comments from social media.

Microblog Retrieval

Interest in microblog retrieval has significantly increased in recent years. Several studies have investigated the nature of microblog search compared to other search tasks. Naveed et al. (2011) illustrated the challenges of microblog retrieval, where documents are very short and typically focus on a single topic. Teevan et al. (2011) highlighted the differences between web queries and microblog queries, where microblog queries usually represent users' interest to find microblog updates about a given event or person as opposed to finding relevant pages on a given topic in web search.

Due to this increased interest in microblog search, TREC introduced a new track focused on microblog retrieval starting from 2011 (Ounis et al., 2011; Soboroff et al., 2012). The track aim was to find the best methods for achieving high precision retrieval of microblogs. A collection of 14 million tweets from Twitter and a test set of 50 topics were provided for investigation (Ounis et al., 2011). Although the track led to a variety of effective retrieval approaches, the issue of modeling the search scenario remains important as the TREC track setup models search like a standard ad-hoc retrieval task, which may be suboptimal (Ounis et al., 2011; Soboroff et al., 2012; Magdy et al., 2012).

Retrieval of microblogs can be a method for obtaining relevant tweets to news articles, which can be classified later into comment/non-comment tweets. However, the state-of-the-art in microblog retrieval and filtering remains insufficient for practical use, where precision is around 0.5 (Ounis et al., 2011; Soboroff et al., 2012).

Comments on News on Twitter

Definition

A comment-tweet for a given piece of news is *a tweet containing information about a user's response towards a news item*. Comment-tweets may express an opinion, an explicit or implicit sentiment, a question, a rumor, personal experience, or a call to action. Comments are not restating or rephrasing news. A comment can be a full sentence describing a user reaction toward the news, an emoticon, or hashtag (#tag). Typically,a tweet contains the headline or a sentence from the news article and often supplies a link to the article.

Comment-tweets vs. Subjective-tweets

Classifying tweets into subjective and objective tweets was investigated in many research studies (Barbosa and Feng, 2010; Chen et al., 2012; Jiang et al., 2011; Lin et al., 2011; Zhao et al., 2011). The common definition of the subjective/sentiment tweets is those that carry opinion/sentiment rather than facts. Identifying subjective tweets is often an initial step in sentiment analysis (Barbosa and Feng, 2010; Kilner and Hoadley, 2005; Wilson et al., 2005). Though some comment-tweets can be subjective, comments need not be subjective.

The main differences between comment-tweets and subjective ones are as follows:

1. Comment-tweets are typically composed of two parts: an objective part that states (or rephrases) the news headline and/or provides a URL; and a commentary part that has the user's comment.
2. Comment-tweet may not be subjective (containing sentiment or opinion). Non-subjective comments may include a call to action (Table 2 ex. 4), an initiation of a discussion (Table 2 ex. 5), or even a correction of the news itself.

Types of Comment-Tweets

People have varying motivations for commenting on news sites (Kilner and Hoadley, 2005). This also applies to comments on news and social sites. Thus, a comment can be expressed in many different ways. We illustrate the possible types of comment-tweets and demonstrate the possible challenges associated with identifying comment-tweets automatically. Table 2 shows some examples of topics along with relevant comment-tweets and their types. In all, Table 2 demonstrates nine different kinds. Some express sentiment toward the news in a sentence or just a word. Others make sarcastic comments. Different types of comments that are expressed in a non-subjective manner include calls to action, discussions, wishes, experiences sharing, and pointers to related articles or blogs. These wide variations of comment types make comment identification challenging. For example, tweets may restate

Table 2. Examples to comments on news from Twitter

#	News Headline	Tweet	Type
1	Drug war comes to Mexico's 2nd city	LOL! @HuffPostWorld Pot-firing catapult found at Mexican border http://huff.to/gCa8sF	Expression of sentiment (short)
2	Phone-hacking in Britain	Funny Guardian thinks Assange is a hero for Wikileaks& NOTW journalists evil for alledged phone hacks. Just pure snobbery &self interest!	Expression of sentiment (long)
3	BBC World Service cuts outlined to staff	If the BBC mortgaged the #Strictly wardrobe to that (minted) wedding dressmaker in #MyBigFatGypsyWedding, they could save the World Service.	Funny/sarcastic
4	U.S. Murder Case Threatens Pakistan Ties	Whether in Egypt or Pakistan the US must demand that our citizens are treated fairly, RELEASE OUR DIPLOMAT! RELEASE OUR JOURNALISTS!	Call for action
5	Haiti's former president Jean-Bertrand Aristide vows to return	If #Aristide returned to #Haiti, would it change anything? Would it create democracy?	Initiate discussion
6	Rachel Maddow at MSNBC makes an idiot of herself again	wish i could afford HBO so i could watch bill maher, and rachel maddow, they tell it like it is.	Wishing/hoping
7	U.S. Unemployment Falls, But New Jobs Lag	First Thoughts: What has changed (and what hasn't): It's bad enough that the progressive income tax, a concept... http://twurl.nl/xzone	Pointer to related blog
8	Al Gore Explains 'Snowmageddon'	Awesome business article related to global warming - crazy stuff - and I'm not a tree hugger! read.bi/gV3WVv via @businessinsider	Pointer to related article
9	TSA shuts door on private airport screening program	Got my first TSA pat down at the Thunder Bay airport. He was friendly	Experience sharing

someone else's message or point to a blog post, such as example 7 in Table 2. Sometimes a link to a blog may be a comment, while other times it might not.

Detecting Comment-Tweets

As we showed, comment-tweets have a wider definition than subjective ones. Relying on term occurrences can be suboptimal or may require large amounts of training data (Lin et al., 2011; Wang et al., 2012). We extracted multiple features from tweets and news articles to train a single SVM model that is capable of classifying comment-tweets of different types. Aside from features that are mentioned in the literature (Luo et al., 2012; Sriram et al., 2010), we introduce here new features. The features that we extracted can be categorized into four groups:

1. **Tweets-specific features (TS):** We used 8 tweet-specific features that relate to how a tweet is written. They are: (1) presence of hashtag (#tag); (2a) presence of user mention (@some_user); (2b) position of the user mention; (3a) presence of link; (3b) position of the link; (4a) presence of "RT", indicating retweet; (4b) the position of "RT"; (5) presence of incomplete text indicated by "…". These features might indicate the presence of a comment in the tweet. For example, the appearance of "RT" in the middle of a tweet probably indicates that a user added some text to the original tweet, probably a comment. Similarly, text after a link may indicate a comment on the linked news item.
2. **Language-independent features (LI)**: We used 7 binary features that indicate the presence of non-lexical markers that may indicate a comment. The features arethe presence of: (1) question marks (?); (2) exclamation marks (!); (3) underscores (_); (4) repeated punctuation marks (e.g. "??" and "!!!!"); (5) emoticons (e.g. ":)", ":(", ":D" … etc.); (6) uppercased words; and (7) elongated words (e.g. "cooool"). These features typically express sentiment and/or commentary.
3. **Lexical features (LX)**: We used 7 binary lexical features that indicate if a tweet contains:
   a. A singular 1st person pronoun (ex. I, me, my, mine).
   b. A question word (e.g. what, why, how).
   c. Sentiment words from the MPQA[2] word list (Riloff and Wiebe, 2003; Wilson et al., 2005). We used the positive, negative, and neutral words to generate three different features.
   d. Social media abbreviations. We used a list of 1,356 abbreviations, such as BRB (be right back), CU (see you), and FYI (for your information). We obtained the abbreviations from Webopedia.com[3]. Some ambiguous abbreviations include: 182 (I hate you), ARE (acronym rich environment), and SO (significant other). Thus, we pruned the list to remove all abbreviations that contained digits only or were among the top 10% most frequent English words (based on the Aspell dictionary[4]). The pruned list contained 1,298 abbreviations.
   e. Expressive words from the dailywritingtips.com[5]. The list of 100 terms contains words such as: boo, ew, ha-ha, uh, and yay.
4. **Topic-dependent features (TD)**. To help identify tweets that paraphrase the news, these features attempt to capture the relevance between tweets and news article. We used four real-valued features that relied on the cosine similarity between a tweet and a news item.

[2] File: subjclueslen1-HLTEMNLP05.tff (http://www.cs.pitt.edu/mpqa/)
[3] http://www.webopedia.com/quick_ref/textmessageabbreviations.asp
[4] http://aspell.net
[5] http://www.dailywritingtips.com/100-mostly-small-but-expressive-interjections/

Cosine similarity was computed between the tweet and either the headline or the headline+body of the news article. Term weights in the feature vector were either the TF or the TF-IDF of the terms.

These sets of features were used to train an SVM classifier.

Experimental Setup

Evaluation of our proposed task required a test set that is composed of a set of news articles and a set of relevant tweets foreach of these articles. Several methods are available for obtaining relevant tweets for an article.The two we experimented with in this paper entail: a) finding tweets that contain a URL pointing to the article; and b) performing a search against tweets using snippets from the article. Other methods may include the use of filtering and topic modeling at article level or at tweet level.

After finding relevant tweets, they are annotated as either comments or not. In this section we describe the construction of the test set and evaluation methodology.

Test Set

We augmented the TREC-2011 Microblog track dataset to build our test set (Ounis et al., 2011). The TREC dataset consisted of approximately 14 million tweets crawled in the period between January 25 and February 8, 2011, inclusive. The tweets collection contained tweets in multiple languages. A set of 50 English topics was provided to the participants in the track, and only English tweets were considered relevant to any of the topics. The topics were expressed using short queries that were typically a few words long. Each topic was associated with a query-time, which is the time of querying this topic on twitter, and the task was to find relevant tweets that were posted before the given time only. Figure 1 shows an example of a microblog track topic. Relevance judgments were constructed by manually assessing the pooled results from the participating systems in the track. The evaluation metric used for evaluation was precision at 30 (P@30), which was picked based on the assumption that users usually checks no more than 30 tweets per query. The number of relevant tweets per topic ranged from zero (only for topic 50, which was excluded) and 200. The relevance assessments of the 49 other topics contained judgments for more than 40k tweets, out of which nearly 3,000 were relevant. Since only English tweets were of interest, we used a language detection tool[6] for filtering non-English tweets. This led to a collection of roughly 5 million English tweets.

Our objective for preparing the test set is to have a set of news articles representing topics instead of the microblog track short queries. For each of the 49 topics, we used the TREC query to search online using Google for relevant news articles in the period between January 25 and topic query-times. A relevant article was manually selected for each of the topics. If no news article was found, we selected the most relevant webpage to the topic instead. Figure 2 shows an example articles for topic "3".

```
<num> Number: MB003 </num>
<title> Haiti Aristide return </title>
<querytime>Feb 08 21:32:13 2011</querytime>
```

Figure 1. Topic "3" in the TREC-2011 microblog track

```
<title>Haiti Aristide return</title>
<source>Guardian</source>
<link>http://www.guardian.co.uk/world/2011/jan/28
/us-blocking-aristide-return-to-haiti</link>
<querytime>Feb 08 21:32:13 2011</querytime>
<articledate>28 January 2011</articledate>
<headline>Haiti's former president Jean-Bertrand
Aristide vows to return</headline>
<subheadline>Ex-leader writes in the Guardian
that his seven-year exile is at an end
</subheadline>
<article>
WikiLeaks confirms what grassroots people have
been saying, which...
</article>
```

Figure 2. News Article for Topic "3"

Despite our best effort, the topical coverage of the selected news articles was not always exactly the same as the original queries of topics. For example, the query for topic "2" was "*2022 FIFA Soccer"*. The selected article was entitled "*Qatar's 2022 FIFA World Cup Stadiums are Eco Marvels*," which is a subtopic of the wider TREC topic. This occurred for some of the topics in the test set, requiring us to reevaluate the relevance assessments to validate that the relevant tweets to the TREC microblog topics were relevant to the selected article. Thus, we made two additional annotations for each relevant tweet to indicate if a tweet is relevant and if it is a comment.

Relevance Assessment Reevaluation

We manually reevaluated all the relevance judgments of all the relevant tweets against the selected articles. Of the original 3,000 relevant tweets for the original topics, we deemed 600 as not relevant to the selected news article.

Moreover, we applied additional retrieval runs for searching the tweets collection using the articles to enrich the test set with additional relevant tweets that may not have been captured in the relevance assessments prepared by the TREC track. We believe that this step was essential since all the assessed tweets by the track organizers were only those retrieved by different participants using the original topics (Ounis et al., 2011).

We indexed the English tweets collection using the Indri toolkit (Strohman et al., 2004). Queries were prepared from the 49 articles using the article headline and sub-headlines (if available). We performed four runs to search the collection:

- **HL**: news article headline as queries
- **HLS**: article headline and sub-headline as queries

[6]http://code.google.com/p/language-detection/

- **HLFB**: similar to HL withpseudo relevance feedback (PRF) using top 50 tweets and 10 terms for the feedback process
- **HLSFB**: similar to HLS + PRF

The top 30 retrieved results from each of the runs were merged and manually judged. An average of 18 tweets per topic were not assessed in the TREC microblog track and required manual assessment. Our assessment led to the addition of 347 relevant tweets to the existing relevance assessments. The final number of relevant tweets for the 49 news articles was 2,890.

Table 3 reports the P@30 and MAP scores for each of the retrieval runs using the new relevance assessments set. These results represent the case where parts of the articles were used to find an initial set of relevant tweets. The average scores achieved compare to the state-of-the-art in microblog search in general (Ounis et al., 2011;Soboroff et al., 2012). For the remainder of the paper, we focus on classifying the relevant tweets as comment-tweets or not.

Comment-Tweets Annotation

We manually assessed the set of relevant tweets to determine if they were comment-tweets or not. The 49 topics were divided among three annotators to manually tag the relevant tweets that represent a user comment on the news article. The annotators were supplied with clear guidelines for tagging the comments. A randomly selected set of 200 tweets out of the full test set were provided to all three annotators in order to calculate the inter-agreement among annotators. The three annotators provided identical annotations for 147 tweets (out of 200). This shows an inter-agreement of 73.5%, which corresponds to a Fliess Kappa of 0.527 indicating moderate agreement. Disagreement was due to some challenging tweets. We also asked the annotators to discuss among themselves doubtful cases they may have encountered and to take a collective decision. Two examples of doubtful tweets that the annotators discussed are shown in Figure 3. The first example shows a tweet that looks like a comment by a user, but it adds nothing to the news. The second tweet reports the news, and the end part of it has a comment by the user that expresses an opinion.

In all, 607 of the relevant tweets were tagged as comment-tweets, which represent roughly 20% of the relevant tweets. Eight of the news articles did not have any corresponding comment-tweets. These eight articles were useful to test the situation when no comments could be identified for a given article. Figure 4 plots the number of relevant tweets for each news article sorted in an ascending order. The portion of the tweets, which were tagged as comments, is indicated in gray.

Table 3.Microblog retrieval results using articles

	HL	HLFB	HLS	HLSFB
P@30	0.443	0.450	0.463	0.472
MAP	0.440	0.452	0.477	0.484

Headline: Taco Bell Sued Over Meat That's Just 35 Percent Beef
Tweet: That ain't necessarily “beef” in your Taco Bell burrito...a new lawsuit wants the chain to label it “taco meat filling”
Final decision: Not a comment

Headline: Rachel Maddow at MSNBC makes an idiot of herself again
Tweet: rachelmaddow is complaining john boehner and jimjordanhave the wrong priorities. but she's running old teletubby
Final decision: A comment

Figure 3 . Examples of doubtful tweets

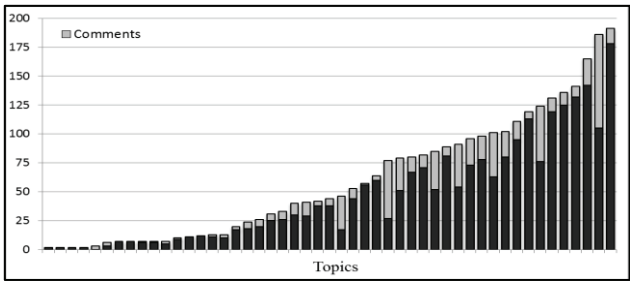

Figure 4 . Number of relevant tweets for the 49 news articles

Evaluation Methodology

Unlike prior work on subjectivity and sentiment analysis of tweets that used accuracy as their measure of goodness, we opted to use precision and recall. Accuracy would have been adequate if the positive and negative classes were comparable in size. This was not the case for comments.

For evaluation, we calculate both precision and recall using micro- and macro-averaging:

1. Micro-average: recall and precision were calculated over identified comment-tweets as a whole without considering which tweet related to which topic. These indicate performance at tweet level.
2. Macro-average: recall and precision were calculated for each topic separately, and then the mean is calculated over all topics. These indicate performance at news article level regardless to the number of relevant tweets.

When using macro-average, there were some situations where some articles did not have any relevant comment-tweets or none were identified. In such cases, either*True Positives* (TP) and *False Negatives* (FN) equaled zero or *TP* and *FP* equaled zero respectively. To overcome these cases, we calculated precision and recall for each topic *t* as:

$$P(t) = \begin{cases} \dfrac{TP(t)}{TP(t) + FP(t)} & TP(t) + FP(t) > 0 \\ 1 & TP(t) = FP(t) = 0 \end{cases}$$

$$R(t) = \begin{cases} \dfrac{TP(t)}{TP(t)+FN(t)} & TP(t)+FN(t) > 0 \\ 1 & TP(t) = FN(t) = 0 \end{cases}$$

Where *FN(t)* is the number of false negatives for topic *t*. When no comments are identified, then precision equals to one. If at least one or more comments exist, then recall would be equal to zero. Otherwise recall would equal one.

Due to the precision oriented nature of the problem, we promoted the importance of precision by using $F_{0.5}$ measure to combine the recall and precision into one score. Showing non-comment tweets to readers is generally undesirable, even if relevant, since they would not add any additional information to the reader.

Experimental Setup

Due to the limited number of training examples, we used cross-validation for training and testing our comment-tweets classification approach (Kohavi, 1995). We specifically applied leave-one-out cross-validation (LOOCV) where one topic was left out and the classifier was trained on the remaining topics.

To evaluate the features set, we tested the classification using different combinations of the four feature groups as follows:

Run1: TS (tweets-specific features)
Run2: LI (language-independent features)
Run3: LX (lexical features)
Run4: TS+LI
Run5: TS+LX
Run6: LI+LX
Run7: TS+LI+LX
Run8: TS+LI+LX+TD (all features including topic-dependent features)

Our aim behind the runs was to understand the effect of each of the feature groups on classification effectiveness.

Using subjectivity analysis for Comment Tweet Classification

In this section we examine the effectiveness of using subjectivity detection as a stand-in for comment-tweets classification. Our examination will lucidly show the differences between subjective and comment tweets. The assumption here is that there is some overlap between comment-tweet detection and subjectivity analysis. For our experiment, we used SentiStrength[7](Thelwall et al., 2010), which is considered a state-of-the-art tool for sentiment analysis for short social text, especially tweets. According to many research studies in subjective/objective classification, subjective text is text that contains polarity of positive or negative sentiment (Barbosa and Feng, 2010; Kilner and Hoadley, 2005; Wilson et al., 2005). Given SentiStrength output, we considered tweets with weak or no sentiment as objective (or not comment-tweets) and those with stronger sentiment as subjective (or comment-tweets). The SentiStrength tool has two modes of classification. The first is the trinary mode, which classifies text into positive/negative/neutral. The second is the scaled mode, which gives a value between -4 to 4 representing extremely negative to extremely positive sentiment respectively. We noticed that the trinary mode classified most tweets as subjective with either positive or negative sentiment. Therefore, we also used classification using the scaled mode, while using different values of sentiment as the threshold for considering if a tweet is subjective or objective. Table 4 reports on our experiments.

Table 4 reports the results of using subjectivity analysis for comment-tweets classification for different scales. As shown in the table, considering tweets with strong sentiment as subjective leads to higher precision, but significantly lower recall. Conversely, considering tweets with lower sentiment as subjective yields higher recall and lower precision. This indicates some correlation between subjectivity analysis and comment-tweet detection. However, the highest achieved micro- and macro-average $F_{0.5}$ values were 0.309 and 0.517 respectively. As we show later, these results are suboptimal.

Table 5 shows some of the examples of non-comment tweets that are classified by the SentiStrength to be subjective tweets, and other examples of comment-tweets that are classified as neutral with no sentiment. These examples highlight again the divergence between subjective and comment-tweets, which further motivates the need for specifically detecting comments.

System Performance

Classification Results on the Test Set

Table 6 reports on the micro- and macro-averaged results of classification using the aforementioned combinations of feature groups. As shown in the table, tweet-specific (TS) features yielded the lowest $F_{0.5}$ values (both micro and macro), compared to other feature combinations. Yet, it still outperformed using subjectivity analysis as a stand-in for comment-tweet detection. Language-independent (LI) features led to higher recall compared to TS with slightly lower precision. Using TS+LI features did not lead to improvements over LI features alone. The lexical (LX) features improved precision over TS and LI, but recall was low. Again, using TS+LX features did not lead to improvement over using LX alone. Combining LI+LX features improved recall (0.573), but with lower precision (0.653) compared to LX. Using all topic-independent features (TS+LI+LX) improved precision and recall

[7]http://sentistrength.wlv.ac.uk/

Table 4.Results of applying subjectivity classification for identifying comment-tweets

Runs	Subjective/Objective based on Sentiment Strength		Macro-average			Micro-average		
	Subjective tweets	Objective tweets	Precision	Recall	$F_{0.5}$	Precision	Recall	$F_{0.5}$
Trinary	Classified tweets as positive or negative	Classified tweets as neutral	0.314	0.722	0.354	0.268	0.810	**0.309**
Scale:1	absolute(Sentiment) ≥ 1	Sentiment = 0	0.305	0.631	0.340	0.254	0.691	0.290
Scale:2	absolute(Sentiment) ≥ 2	-2 < Sentiment < 2	0.520	0.390	0.488	0.271	0.284	0.273
Scale:3	absolute(Sentiment) ≥ 3	-3 < Sentiment < 3	0.752	0.230	**0.517**	0.214	0.082	0.162
Scale:4	absolute(Sentiment) = 4	-4 < Sentiment < 4	0.939	0.167	0.488	0.571	0.007	0.032

Table 5. Examples of false positive and false negative classifications using a subjectivity analysis for comment-tweet detection

	News Headline	**Tweet**	**Sentiment Strength**
Non-comment-tweets classified subjective	Drug war comes to Mexico's 2nd city	Horribly Mutilated Bodies Discovered In Mexico : huffingtonpost.com/2011/01/31/mexâ€	-3
	U.S. Murder Case Threatens Pakistan Ties	DTN Pakistan: Faheem"s wife death "tragedy": US: WASHINGTON: The United States on Monday called the suicide ... http://bit.ly/gihnDx	-3
Comment-tweets classified not subjective	Haiti's former president Jean-Bertrand Aristide vows to return	If #Aristide returned to #Haiti, would it change anything? Would it create democracy?	0
	Egyptians form human shield to protect museum	Hopefully, someone will remember that the NDP headquarters in Cairo are next to the Cairo Museum. #egypt	0

Table 6. Classification results for comment-tweet detection for 8 runs of different combinations of features

	Macro-average			Micro-average		
	P	R	$F_{0.5}$	P	R	$F_{0.5}$
TS	0.691	0.323	0.563	0.582	0.190	0.412
LI	0.672	0.470	0.619	0.670	0.407	0.593
LX	0.789	0.350	0.631	0.738	0.297	0.569
TS+LI	0.648	0.472	0.603	0.646	0.412	0.580
TS+LX	0.736	0.384	0.622	0.720	0.363	0.602
LI+LX	0.653	0.573	0.635	0.663	**0.558**	0.639
TS+LI+LX	0.809	**0.578**	0.749	0.795	0.528	**0.722**
All Feats	**0.884**	0.503	**0.768**	**0.843**	0.450	0.718

compared to using any of them alone. Adding the topic-dependent (TD) features to the set of features led to a significant improvement in precision to reach 0.884, but a drop in the recall to 0.503. Using all features led to the best overall macro-average $F_{0.5}$ and nearly the highest micro-average $F_{0.5}$. The difference in results for using all features compared to using topic-independent features only was statistically significant. We used a paired 2-tailed *t*-test with *p*-value less than 0.05 to test statistical significance.

Based on the results in Table 6, the best configuration for classification is achieved when either using all the topic independent features alone or in combination with TD. Although the best achieved recall was 0.50-0.60, this task is precision oriented and the large number of tweets can compensate for low recall. Precision is relatively high with values of over 0.84. In practice, 8 to 9 out of 10 suggested comment-tweets are in fact comment-tweets.

Analysis

We analyzed the output by examining the false negatives, which affect recall, and false positives, which affect precision. Concerning false negatives, the main problems that we identified were as follows:

1. Lack of features that would lead to a correct classification:
   a. The comment tweet is so short that it is difficult to extract useful features from it. E.g. "*the rite wasn't even all that*".
   b. In a longer tweet, the manner of expression does not offer many features to be extracted. E.g. "*It's not all bad news for jobs. Unemployment rate fell from 9.4% to 9%. Avg. hourly earnings up 0.4%. Nov. & Dec job growth revised higher.*" Only two features were present: the sentiment word "bad" and the beginning with a capital letter "It".
2. The limited coverage of the lexicons that we used. For example, the lexicons do not cover some curse words such as "*bloody*", "*suck*" … etc.
3. Though the presence of emoticons, elongated words, and "emotion" words may be strong indicators, they did not occur frequently enough in the training corpus.

As for false positives, we identified the following reasons for misclassification:

1. A tweet that contains a question whose answer is in the article. Consider the tweet: "*Why is it so cold, if global warming is such a big deal? http://bit.ly/hbFv8X*", where the link refers to an article on global warming. Here the question does not represent any response from the user, and hence it is not a comment.

2. A tweet contains the headline of a relevant article to the news article that is in the form of a comment. Consider the example tweet: "*Maddow's Excuse for Reporting Spoof Story as Fact: It's Beck's Fault! http://bit.ly/euNxOv*". This kind of problem can be partially resolved using topic-dependent features. However, when a different article rephrases the news in a comment-like way, it becomes difficult for our classifier to properly classify it.

We think that the classifier can benefit from more training examples to resolve most of these issues. Nonetheless, the classifier that we trained can effectively identify comment tweets, and such tweets would likely improve readers' experience

System Validation on Live News Articles

After testing comment classification system on our test set, we validated its performance on live news articles. We selected 15 articles on popular news topics that were published between the end of July and beginning of August 2012 to test our system. We collected the articles from different popular news websites, such as CNN, NY-Times, the Economist, Reuters, and Al-Jazeera. To ensure the retrieval of relevant tweets for the news articles, we used the article URL to search Twitter for tweets linking to the article. This guaranteed that the retrieved tweets are most likely relevant to the linked news article. We collected the most recent 100 tweets linking the article. In all, we scraped a total of 1,384 tweets for the 15 articles, where some articles were tweeted less than 100 times.

The 49 articles from the aforementioned dataset were used for training the classifier. Out of the 1,384 tweets, only 99 tweets were classified as comments. Since this is a precision oriented task, we evaluated classification performance using precision. We found that among the 99 tweets classified as comment, 95 were indeed comment-tweets. All false positive comment-tweets were associated with the same article.

Table 7 shows some examples of the validation new articles and samples of the corresponding identified comment-tweets. The last example in Table 7 is the one that had four misclassified comment-tweets (the tweet in italic format). The four were retweets of the same, hence they were identical. The tweet rephrased the headline of the article and is not a comment.

It is noteworthy that the number of retweets of a tweet can indicate its importance. Consider the first identified comment tweet for the first example article in Table 7: "*What do #India and #Pakistan have in common? The need for effective regulation of their respective electrical grid. http://t.co/8I0jbQwE*". This tweet was retweeted seven times, and the initialtweet was authored by "Philip J. Crowley", who describes himself on Twitter as a "Fellow at The George Washington University Institute for Public Diplomacy and Global Communication and Commentator for the BBC and Daily Beast" and has more than 52,500 followers. Promoting comment-tweets by the number of retweets or the number of followers of the original author can further improve a reader's experience. Such features/rankings are generally not available in commentary frameworks on news website. In future work, we would like to investigate ways to glean comment-tweets experts or

Table 7. The headlines of news articles, the number of retrieved tweets for each article from Twitter, the number of classified tweets as comments by our system, and sample of the identified comments for each news article.

News Headline	Ret	Idnt	Samples of the identified comments
2nd Day of Power Failures Cripples Wide Swath of India	99	20	• What do #India and #Pakistan have in common? The need for effective regulation of their respective electrical grid. http://t.co/8I0jbQwE • 2nd Day of Power Failures Cripples Wide Swath of India http://t.co/Pd1s14e1 If Obama policies continue, it can & will happen in the US??
Mr. Bean Gets Carried Away During Olympics Appearance	98	6	• What I like about Olympics. :))) http://t.co/SnLB6XYb • Funny! LOL! http://t.co/lvKOz3p8
Reports: iPhone 5 to be unveiled Sept. 12	54	9	• @ClaudeBotes87 - on my bday?? Its a sign! http://t.co/1Dj0YUn2 • Give me that!!! Reports: iPhone 5 to be unveiled 9/12. http://t.co/fWdtAg0D #cnn #apple
Fight continues for control of Syria's Aleppo	100	1	• Fight continues for control of Syria's Aleppo - http://t.co/LsJ1p3gh Call in IZZY for counter air support!!!
Obama announces new Iran sanctions	99	4	• Obama announces new Iran sanctions http://t.co/n1Y5d4TY <-how many more are possible? I hear about new Iran sanctions every other week. • @AJEnglish: Obama announces new Iran sanctions http://t.co/nZiHjqgh - who's the real evil?
Peter Jackson's The Hobbit to be extended to three films	100	24	• Oh lord, now the Hobbit is going to be a trilogy. Honestly, that man doesn't know how to edit. http://t.co/cTDONkGT • hmm ... Artistic vision or studio money grabbing? 3 films for the price of 3, instead of 2, or just be concise and do 1 http://t.co/FyZiRB2g
London 2012: rowers Glover and Stanning win Team GB's first gold medal	55	5	• Gold Medal commemorative stamps http://t.co/d2iBsU7T available from Post Offices from tomorrow. Go Team GB! http://t.co/YtDiUDWi • *It's gold! Stanning and Glover end the wait for Team GB http://t.co/xrdi20jm #teamfollowback*

celebrities on their own social networks even if they don't directly comment on the article of interest.

The samples of the live articles presented in this section demonstrate the effectiveness of our classification features in identifying comment-tweets for a given article in a practical environment. In this, we used tweets linking to articles, which we assumed are relevant. Using state-of-the-art microblog retrieval yields relatively low precision (less 0.50) (Ounis et al., 2011; Soboroff et al., 2012). Perhaps improved retrieval of tweets that relate to articles or interactive filtering may be required to yield a more precise set of relevant tweets prior to classifying them as comments or not.

Conclusion and Future Work

In this paper, we proposed the novel task of detecting comment-tweets on news articles from Twitter. Offering such comment-tweets alongside news articles can enrich readers' experience. We defined comment-tweets, and we contrasted them against subjective-tweets. We also described a supervised learning approach using SVM classification to properly identify such tweets. We explained the set of features we used in detail, and tested the effectiveness of each. We built a test set for evaluating the task and described the evaluation methodology. Our experimental results showed our classification's high performance for detecting comment-tweets with high precision. We provided error-analysis to identify problems that led to false positives and false negatives. Finally, we demonstrated the high performance of our system on live news articles and how popularity of tweets can potentially be used to rank comment-tweets.

The work presented in this paper opens a new research direction for social text classification, which we shown to have a practical use and benefits to users. Although we achieved relatively high precision, recall still requires improvement. We will provide all our annotated data and resources of classification online for researchers who are interested in the task. An essential component that will maximize the benefit of our classifier is a high precision microblog filtering/retrieval system. The precision of state-of-the-art microblog retrieval is almost 50%. Using tweets linking to articles seems like a reasonable choice until filtering/retrieval can be more precise. Finally, we think that building a dedicated classifier for each type of comment shown in Table 2 can lead to improved results.

References

Barbosa L., J. Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. *COLING 2010.*

Chen L., W. Wang, M. Nagarajan, S. Wang, A. P. Sheth. 2012. Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter. *ICWSM 2012.*

Diakopoulos N., M. Naaman. 2011. Towards Quality Discourse in Online News Comments. *CSCW 2011.*

Gao W., P. Li, K. Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. *CIKM 2012*

Gilbert E., K. Karahalios. 2010. Understanding Deja Reviewers. *CSCW 2010.*

González-Ibáñez R., S.Muresan, N. Wacholder.2011. Identifying Sarcasm in Twitter: A Closer Look. *ACL 2011.*

Kilner P., C. Hoadley. 2005. Anonymity Options and Professional Participation in an Online Community of Practice. *CSCL 2005.*

Kohavi R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection.*IJCAI 1995.*

Jiang L., M. Yu, M. Zhou, X. Liu, T. Zhao. 2011. Target-dependent Twitter Sentiment Classification *ACL 2011.*

Lin C., Y. He, R. Everson. 2011. Sentence Subjectivity Detection with Weakly-Supervised Learning. *IJCNLP 2011.*

Liu B. 2009. Sentiment analysis and subjectivity.*Handbook of Natural Language Processing*, 627-666.

Luo Z., M. Osborne, T. Wang. 2012. Opinion Retrieval in Twitter. *ICWSM 2012.*

Magdy W., A. Ali, K. Darwish. 2012. A Summarization Tool for Time-Sensitive Social Media. *CIKM 2012.*

Naveed N., T. Gottron, J. Kunegis, A. Alhadi. 2011. Searching microblogs: coping with sparsity and document quality. *CIKM 2011*.

Ounis I., C. Macdonald, J. Lin, I. Soboroff. 2011. Overview of the TREC-2011 Microblog Track. *TREC 2011.*

Phelan O., K. McCarthy, M. Bennett, and B. Smyth. 2011. Terms of a feather: content-based news recommendation and discovery using twitter. *ECIR 2011*.

Riloff E. and T. Wiebe 2003. Learning extraction patterns for subjective expressions.*EMNLP 2003.*

Soboroff I., I. Ounis, J. Lin, I. Soboroff. 2012. Overview of the TREC-2012 Microblog Track. *TREC 2012.*

Sriram B., D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas. 2010. Short Text Classification in Twitter to Improve Information Filtering.*SIGIR 2010.*

Strohman T., D. Metzler, H. Turtle, W. B. Croft. 2004. Indri: A language model-based search engine for complex queries. *ICIA 2004.*

Subasic I., B. Berendt. 2011. Peddling or Creating? Investigating the Role of Twitter in News Reporting.*ECIR-2011.*

Teevan J., D. Ramage, M. Morris. 2011. #Twittersearch: A comparison of microblog search and web search. *WSDM 2011*.

Thelwall M., K. Buckley, G. Paltoglou, D. Cai, A. Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.*

Wang H., D. Can, A. Kazemzadeh, F. Bar, S. Narayanan. 2012. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. *ACL 2012.*

Wilson T., J.Wiebe and P. Hoffmann 2005.Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.*HLT 2005.*

Yang Z., K. Caiy, J. Tang, L. Zhangy, Z. Suy.and J. Li. 2011. SocialContext Summarization. *SIGIR 2011.*

Zhao W. X., J. Jiang, Ji. Weng, J.He, E-P. Lim, Ho. Yan, X. Li. 2011. Comparing twitter and traditional media using topic models. *ECIR 2011*.