Target-dependent Twitter Sentiment Classification

Long Jiang[1] **Mo Yu**[2] **Ming Zhou**[1] **Xiaohua Liu**[1] **Tiejun Zhao**[2]

1 Microsoft Research Asia
Beijing, China
{longj,mingzhou,xiaoliu}@microsoft.com

2 School of Computer Science & Technology
Harbin Institute of Technology
Harbin, China
{yumo,tjzhao}@mtlab.hit.edu.cn

Abstract

Sentiment analysis on Twitter data has attracted much attention recently. In this paper, we focus on target-dependent Twitter sentiment classification; namely, given a query, we classify the sentiments of the tweets as positive, negative or neutral according to whether they contain positive, negative or neutral sentiments about that query. Here the query serves as the target of the sentiments. The state-of-the-art approaches for solving this problem always adopt the target-independent strategy, which may assign irrelevant sentiments to the given target. Moreover, the state-of-the-art approaches only take the tweet to be classified into consideration when classifying the sentiment; they ignore its context (i.e., related tweets). However, because tweets are usually short and more ambiguous, sometimes it is not enough to consider only the current tweet for sentiment classification. In this paper, we propose to improve target-dependent Twitter sentiment classification by 1) incorporating target-dependent features; and 2) taking related tweets into consideration. According to the experimental results, our approach greatly improves the performance of target-dependent sentiment classification.

1 Introduction

Twitter, as a micro-blogging system, allows users to publish tweets of up to 140 characters in length to tell others what they are doing, what they are thinking, or what is happening around them. Over the past few years, Twitter has become very popular. According to the latest Twitter entry in Wikipedia, the number of Twitter users has climbed to 190 million and the number of tweets published on Twitter every day is over 65 million[1].

As a result of the rapidly increasing number of tweets, mining people's sentiments expressed in tweets has attracted more and more attention. In fact, there are already many web sites built on the Internet providing a Twitter sentiment search service, such as Tweetfeel[2], Twendz[3], and Twitter Sentiment[4]. In those web sites, the user can input a sentiment target as a query, and search for tweets containing positive or negative sentiments towards the target. The problem needing to be addressed can be formally named as Target-dependent Sentiment Classification of Tweets; namely, given a query, classifying the sentiments of the tweets as positive, negative or neutral according to whether they contain positive, negative or neutral sentiments about that query. Here the query serves as the target of the sentiments.

The state-of-the-art approaches for solving this problem, such as (Go et al., 2009[5]; Barbosa and Feng, 2010), basically follow (Pang et al., 2002), who utilize machine learning based classifiers for the sentiment classification of texts. However, their classifiers actually work in a target-independent way: all the features used in the classifiers are independent of the target, so the sentiment is decided no matter what the target is. Since (Pang et al., 2002) (or later research on sentiment classification

[1] http://en.wikipedia.org/wiki/Twitter
[2] http://www.tweetfeel.com/
[3] http://twendz.waggeneredstrom.com/
[4] http://twittersentiment.appspot.com/
[5] The algorithm used in Twitter Sentiment

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 151–160,
Portland, Oregon, June 19-24, 2011.

of product reviews) aim to classify the polarities of movie (or product) reviews and each movie (or product) review is assumed to express sentiments only about the target movie (or product), it is reasonable for them to adopt the target-independent approach. However, for target-dependent sentiment classification of tweets, it is not suitable to exactly adopt that approach. Because people may mention multiple targets in one tweet or comment on a target in a tweet while saying many other unrelated things in the same tweet, target-independent approaches are likely to yield unsatisfactory results:

1. Tweets that do not express any sentiments to the given target but express sentiments to other things will be considered as being opinionated about the target. For example, the following tweet expresses no sentiment to *Bill Gates* but is very likely to be classified as positive about *Bill Gates* by target-independent approaches.

   "People everywhere love Windows & vista. Bill Gates"

2. The polarities of some tweets towards the given target are misclassified because of the interference from sentiments towards other targets in the tweets. For example, the following tweet expresses a positive sentiment to *Windows 7* and a negative sentiment to *Vista*. However, with target-independent sentiment classification, both of the targets would get positive polarity.

   "Windows 7 is much better than Vista!"

In fact, it is easy to find many such cases by looking at the output of Twitter Sentiment or other Twitter sentiment analysis web sites. Based on our manual evaluation of Twitter Sentiment output, about 40% of errors are because of this (see Section 6.1 for more details).

In addition, tweets are usually shorter and more ambiguous than other sentiment data commonly used for sentiment analysis, such as reviews and blogs. Consequently, it is more difficult to classify the sentiment of a tweet only based on its content. For instance, for the following tweet, which contains only three words, it is difficult for any existing approaches to classify its sentiment correctly.

"First game: Lakers!"

However, relations between individual tweets are more common than those in other sentiment data. We can easily find many related tweets of a given tweet, such as the tweets published by the same person, the tweets replying to or replied by the given tweet, and retweets of the given tweet. These related tweets provide rich information about what the given tweet expresses and should definitely be taken into consideration for classifying the sentiment of the given tweet.

In this paper, we propose to improve target-dependent sentiment classification of tweets by using both target-dependent and context-aware approaches. Specifically, the target-dependent approach refers to incorporating syntactic features generated using words syntactically connected with the given target in the tweet to decide whether or not the sentiment is about the given target. For instance, in the second example, using syntactic parsing, we know that "*Windows 7*" is connected to "*better*" by a copula, while "*Vista*" is connected to "*better*" by a preposition. By learning from training data, we can probably predict that "*Windows 7*" should get a positive sentiment and "*Vista*" should get a negative sentiment.

In addition, we also propose to incorporate the contexts of tweets into classification, which we call a context-aware approach. By considering the sentiment labels of the related tweets, we can further boost the performance of the sentiment classification, especially for very short and ambiguous tweets. For example, in the third example we mentioned above, if we find that the previous and following tweets published by the same person are both positive about the *Lakers*, we can confidently classify this tweet as positive.

The remainder of this paper is structured as follows. In Section 2, we briefly summarize related work. Section 3 gives an overview of our approach. We explain the target-dependent and context-aware approaches in detail in Sections 4 and 5 respectively. Experimental results are reported in Section 6 and Section 7 concludes our work.

2 Related Work

In recent years, sentiment analysis (SA) has become a hot topic in the NLP research community. A lot of papers have been published on this topic.

2.1 Target-independent SA

Specifically, Turney (2002) proposes an unsupervised method for classifying product or movie reviews as positive or negative. In this method, sentimental phrases are first selected from the reviews according to predefined part-of-speech patterns. Then the semantic orientation score of each phrase is calculated according to the mutual information values between the phrase and two predefined seed words. Finally, a review is classified based on the average semantic orientation of the sentimental phrases in the review.

In contrast, (Pang et al., 2002) treat the sentiment classification of movie reviews simply as a special case of a topic-based text categorization problem and investigate three classification algorithms: Naive Bayes, Maximum Entropy, and Support Vector Machines. According to the experimental results, machine learning based classifiers outperform the unsupervised approach, where the best performance is achieved by the SVM classifier with unigram presences as features.

2.2 Target-dependent SA

Besides the above mentioned work for target-independent sentiment classification, there are also several approaches proposed for target-dependent classification, such as (Nasukawa and Yi, 2003; Hu and Liu, 2004; Ding and Liu, 2007). (Nasukawa and Yi, 2003) adopt a rule based approach, where rules are created by humans for adjectives, verbs, nouns, and so on. Given a sentiment target and its context, part-of-speech tagging and dependency parsing are first performed on the context. Then predefined rules are matched in the context to determine the sentiment about the target. In (Hu and Liu, 2004), opinions are extracted from product reviews, where the features of the product are considered opinion targets. The sentiment about each target in each sentence of the review is determined based on the dominant orientation of the opinion words appearing in the sentence.

As mentioned in Section 1, target-dependent sentiment classification of review sentences is quite different from that of tweets. In reviews, if any sentiment is expressed in a sentence containing a feature, it is very likely that the sentiment is about the feature. However, the assumption does not hold in tweets.

2.3 SA of Tweets

As Twitter becomes more popular, sentiment analysis on Twitter data becomes more attractive. (Go et al., 2009; Parikh and Movassate, 2009; Barbosa and Feng, 2010; Davidiv et al., 2010) all follow the machine learning based approach for sentiment classification of tweets. Specifically, (Davidiv et al., 2010) propose to classify tweets into multiple sentiment types using hashtags and smileys as labels. In their approach, a supervised KNN-like classifier is used. In contrast, (Barbosa and Feng, 2010) propose a two-step approach to classify the sentiments of tweets using SVM classifiers with abstract features. The training data is collected from the outputs of three existing Twitter sentiment classification web sites. As mentioned above, these approaches work in a target-independent way, and so need to be adapted for target-dependent sentiment classification.

3 Approach Overview

The problem we address in this paper is target-dependent sentiment classification of tweets. So the input of our task is a collection of tweets containing the target and the output is labels assigned to each of the tweets. Inspired by (Barbosa and Feng, 2010; Pang and Lee, 2004), we design a three-step approach in this paper:

1. Subjectivity classification as the first step to decide if the tweet is subjective or neutral about the target;
2. Polarity classification as the second step to decide if the tweet is positive or negative about the target if it is classified as subjective in Step 1;
3. Graph-based optimization as the third step to further boost the performance by taking the related tweets into consideration.

In each of the first two steps, a binary SVM classifier is built to perform the classification. To train the classifiers, we use SVM-Light[6] with a linear kernel; the default setting is adopted in all experiments.

[6] http://svmlight.joachims.org/

3.1 Preprocessing

In our approach, rich feature representations are used to distinguish between sentiments expressed towards different targets. In order to generate such features, much NLP work has to be done beforehand, such as tweet normalization, POS tagging, word stemming, and syntactic parsing.

In our experiments, POS tagging is performed by the OpenNLP POS tagger[7]. Word stemming is performed by using a word stem mapping table consisting of about 20,000 entries. We also built a simple rule-based model for tweet normalization which can correct simple spelling errors and variations into normal form, such as "gooood" to "good" and "luve" to "love". For syntactic parsing we use a Maximum Spanning Tree dependency parser (McDonald et al., 2005).

3.2 Target-independent Features

Previous work (Barbosa and Feng, 2010; Davidiv et al., 2010) has discovered many effective features for sentiment analysis of tweets, such as emoticons, punctuation, prior subjectivity and polarity of a word. In our classifiers, most of these features are also used. Since these features are all generated without considering the target, we call them target-independent features. In both the subjectivity classifier and polarity classifier, the same target-independent feature set is used. Specifically, we use two kinds of target-independent features:

1. Content features, including words, punctuation, emoticons, and hashtags (hashtags are provided by the author to indicate the topic of the tweet).
2. Sentiment lexicon features, indicating how many positive or negative words are included in the tweet according to a predefined lexicon. In our experiments, we use the lexicon downloaded from General Inquirer[8].

4 Target-dependent Sentiment Classification

Besides target-independent features, we also incorporate target-dependent features in both the subjectivity classifier and polarity classifier. We will explain them in detail below.

4.1 Extended Targets

It is quite common that people express their sentiments about a target by commenting not on the target itself but on some related things of the target. For example, one may express a sentiment about a company by commenting on its products or technologies. To express a sentiment about a product, one may choose to comment on the features or functionalities of the product. It is assumed that readers or audiences can clearly infer the sentiment about the target based on those sentiments about the related things. As shown in the tweet below, the author expresses a positive sentiment about "Microsoft" by expressing a positive sentiment directly about "Microsoft technologies".

"I am passionate about Microsoft technologies especially Silverlight."

In this paper, we define those aforementioned related things as Extended Targets. Tweets expressing positive or negative sentiments towards the extended targets are also regarded as positive or negative about the target. Therefore, for target-dependent sentiment classification of tweets, the first thing is identifying all extended targets in the input tweet collection.

In this paper, we first regard all noun phrases, including the target, as extended targets for simplicity. However, it would be interesting to know under what circumstances the sentiment towards the target is truly consistent with that towards its extended targets. For example, a sentiment about someone's behavior usually means a sentiment about the person, while a sentiment about someone's colleague usually has nothing to do with the person. This could be a future work direction for target-dependent sentiment classification.

In addition to the noun phrases including the target, we further expand the extended target set with the following three methods:

1. Adding mentions co-referring to the target as new extended targets. It is common that people use definite or demonstrative noun phrases or pronouns referring to the target in a tweet and express sentiments directly on them. For instance, in "Oh, Jon Stewart. How I love you so.", the author expresses

[7] http://opennlp.sourceforge.net/projects.html
[8] http://www.wjh.harvard.edu/~inquirer/

a positive sentiment to "you" which actually refers to "Jon Stewart". By using a simple co-reference resolution tool adapted from (Soon et al., 2001), we add all the mentions referring to the target into the extended target set.

2. Identifying the top K nouns and noun phrases which have the strongest association with the target. Here, we use Pointwise Mutual Information (PMI) to measure the association.

$$PMI(w,t) = \log \frac{p(w,t)}{p(w)p(t)}$$

Where *p(w,t)*, *p(w)*, and *p(t)* are probabilities of *w* and *t* co-occurring, *w* appearing, and *t* appearing in a tweet respectively. In the experiments, we estimate them on a tweet corpus containing 20 million tweets. We set $K = 20$ in the experiments based on empirical observations.

3. Extracting head nouns of all extended targets, whose PMI values with the target are above some predefined threshold, as new extended targets. For instance, suppose we have found *"Microsoft Technologies"* as the extended target, we will further add *"technologies"* into the extended target set if the PMI value for *"technologies"* and *"Microsoft"* is above the threshold. Similarly, we can find *"price"* as the extended targets for *"iPhone"* from *"the price of iPhone"* and *"LoveGame"* for *"Lady Gaga"* from *"LoveGame by Lady Gaga"*.

4.2 Target-dependent Features

Target-dependent sentiment classification needs to distinguish the expressions describing the target from other expressions. In this paper, we rely on the syntactic parse tree to satisfy this need. Specifically, for any word stem w_i in a tweet which has one of the following relations with the given target *T* or any from the extended target set, we generate corresponding target-dependent features with the following rules:

- w_i is a transitive verb and *T* (or any of the extended target) is its object; we generate a feature w_i *_arg2*. "arg" is short for "argument". For example, for the target iPhone in *"I love iPhone"*, we generate *"love_arg2"* as a feature.
- w_i is a transitive verb and *T* (or any of the extended target) is its subject; we generate a feature w_i*_arg1* similar to Rule 1.
- w_i is a intransitive verb and *T* (or any of the extended target) is its subject; we generate a feature w_i*_it_arg1*.
- w_i is an adjective or noun and *T* (or any of the extended target) is its head; we generate a feature w_i*_arg1*.
- w_i is an adjective or noun and it (or its head) is connected by a copula with *T* (or any of the extended target); we generate a feature w_i*_cp_arg1*.
- w_i is an adjective or intransitive verb appearing alone as a sentence and *T* (or any of the extended target) appears in the previous sentence; we generate a feature w_i*_arg*. For example, in "*John did that. Great!*", "*Great*" appears alone as a sentence, so we generate "*great_arg*" for the target "*John*".
- w_i is an adverb, and the verb it modifies has *T* (or any of the extended target) as its subject; we generate a feature *arg1_v_*w_i. For example, for the target *iPhone* in the tweet "*iPhone works better with the CellBand",* we will generate the feature "*arg1_v_well*".

Moreover, if any word included in the generated target-dependent features is modified by a negation[9], then we will add a prefix "*neg-*" to it in the generated features. For example, for the target iPhone in the tweet "*iPhone does not work better with the CellBand*", we will generate the features "*arg1_v_neg-well*" and "*neg-work_it_arg1*".

To overcome the sparsity of target-dependent features mentioned above, we design a special binary feature indicating whether or not the tweet contains at least one of the above target-dependent features. Target-dependent features are binary features, each of which corresponds to the presence of the feature in the tweet. If the feature is present, the entry will be 1; otherwise it will be 0.

[9] Seven negations are used in the experiments: *not, no, never, n't, neither, seldom, hardly*.

5 Graph-based Sentiment Optimization

As we mentioned in Section 1, since tweets are usually shorter and more ambiguous, it would be useful to take their contexts into consideration when classifying the sentiments. In this paper, we regard the following three kinds of related tweets as context for a tweet.

1. Retweets. Retweeting in Twitter is essentially the forwarding of a previous message. People usually do not change the content of the original tweet when retweeting. So retweets usually have the same sentiment as the original tweets.
2. Tweets containing the target and published by the same person. Intuitively, the tweets published by the same person within a short timeframe should have a consistent sentiment about the same target.
3. Tweets replying to or replied by the tweet to be classified.

Based on these three kinds of relations, we can construct a graph using the input tweet collection of a given target. As illustrated in Figure 1, each circle in the graph indicates a tweet. The three kinds of edges indicate being published by the same person (solid line), retweeting (dash line), and replying relations (round dotted line) respectively.

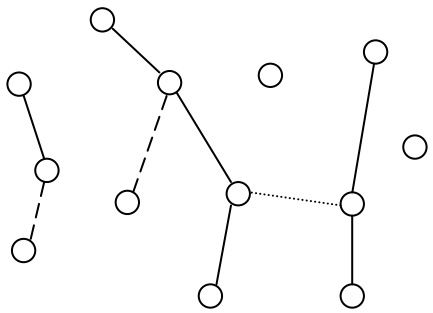

Figure 1. An example graph of tweets about a target

If we consider that the sentiment of a tweet only depends on its content and immediate neighbors, we can leverage a graph-based method for sentiment classification of tweets. Specifically, the probability of a tweet belonging to a specific sentiment class can be computed with the following formula:

$$p(c \mid \tau, G) = p(c \mid \tau) \sum_{N(d)} p(c \mid N(d)) p(N(d))$$

Where c is the sentiment label of a tweet which belongs to {positive, negative, neutral}, G is the tweet graph, $N(d)$ is a specific assignment of sentiment labels to all immediate neighbors of the tweet, and τ is the content of the tweet.

We can convert the output scores of a tweet by the subjectivity and polarity classifiers into probabilistic form and use them to approximate $p(c \mid \tau)$. Then a relaxation labeling algorithm described in (Angelova and Weikum, 2006) can be used on the graph to iteratively estimate $p(c \mid \tau, G)$ for all tweets. After the iteration ends, for any tweet in the graph, the sentiment label that has the maximum $p(c \mid \tau, G)$ is considered the final label.

6 Experiments

Because there is no annotated tweet corpus publicly available for evaluation of target-dependent Twitter sentiment classification, we have to create our own. Since people are most interested in sentiments towards celebrities, companies and products, we selected 5 popular queries of these kinds: {*Obama*, *Google*, *iPad*, *Lakers*, *Lady Gaga*}. For each of those queries, we downloaded 400 English tweets[10] containing the query using the Twitter API.

We manually classify each tweet as positive, negative or neutral towards the query with which it is downloaded. After removing duplicate tweets, we finally obtain 459 positive, 268 negative and 1,212 neutral tweets.

Among the tweets, 100 are labeled by two human annotators for inter-annotator study. The results show that for 86% of them, both annotators gave identical labels. Among the 14 tweets which the two annotators disagree on, only 1 case is a positive-negative disagreement (one annotator considers it positive while the other negative), and the other 13 are all neutral-subjective disagreement. This probably indicates that it is harder for humans to decide if a tweet is neutral or subjective than to decide if it is positive or negative.

[10] In this paper, we use sentiment classification of English tweets as a case study; however, our approach is applicable to other languages as well.

6.1 Error Analysis of Twitter Sentiment Output

We first analyze the output of Twitter Sentiment (TS) using the five test queries. For each query, we randomly select 20 tweets labeled as positive or negative by TS. We also manually classify each tweet as positive, negative or neutral about the corresponding query. Then, we analyze those tweets that get different labels from TS and humans. Finally we find two major types of error: 1) Tweets which are totally neutral (for any target) are classified as subjective by TS; 2) sentiments in some tweets are classified correctly but the sentiments are not truly about the query. The two types take up about 35% and 40% of the total errors, respectively.

The second type is actually what we want to resolve in this paper. After further checking those tweets of the second type, we found that most of them are actually neutral for the target, which means that the dominant error in Twitter Sentiment is classifying neutral tweets as subjective. Below are several examples of the second type where the bolded words are the targets.

*"No debate needed, heat can't beat **lakers** or celtics"* (negative by TS but positive by human)

*"why am i getting spams from weird people asking me if i want to chat with **lady gaga**"* (positive by TS but neutral by human)

*"Bringing iPhone and **iPad** apps into cars? http://www.speakwithme.com/ will be out soon and alpha is awesome in my car."* (positive by TS but neutral by human)

*"Here's a great article about Monte Veronese cheese. It's in Italian so just put the url into **Google** translate and enjoy http://ow.ly/3oQ77"* (positive by TS but neutral by human)

6.2 Evaluation of Subjectivity Classification

We conduct several experiments to evaluate subjectivity classifiers using different features. In the experiments, we consider the positive and negative tweets annotated by humans as subjective tweets (i.e., positive instances in the SVM classifiers), which amount to 727 tweets. Following (Pang et al., 2002), we balance the evaluation data set by randomly selecting 727 tweets from all neutral tweets annotated by humans and consider them as objective tweets (i.e., negative instances in the classifiers). We perform 10-fold cross-validations on the selected data. Following (Go et al., 2009; Pang et al., 2002), we use accuracy as a metric in our experiments. The results are listed below.

Features	Accuracy (%)
Content features	61.1
+ Sentiment lexicon features	63.8
+ Target-dependent features	**68.2**
Re-implementation of (Barbosa and Feng, 2010)	60.3

Table 1. Evaluation of subjectivity classifiers.

As shown in Table 1, the classifier using only the content features achieves an accuracy of 61.1%. Adding sentiment lexicon features improves the accuracy to 63.8%. Finally, the best performance (68.2%) is achieved by combining target-dependent features and other features (t-test: $p < 0.005$). This clearly shows that target-dependent features do help remove many sentiments not truly about the target. We also re-implemented the method proposed in (Barbosa and Feng, 2010) for comparison. From Table 1, we can see that all our systems perform better than (Barbosa and Feng, 2010) on our data set. One possible reason is that (Barbosa and Feng, 2010) use only abstract features while our systems use more lexical features.

To further evaluate the contribution of target extension, we compare the system using the exact target and all extended targets with that using only the exact target. We also eliminate the extended targets generated by each of the three target extension methods and reevaluate the performances.

Target	Accuracy (%)
Exact target	65.6
+ all extended targets	**68.2**
- co-references	68.0
- targets found by PMI	67.8
- head nouns	67.3

Table 2. Evaluation of target extension methods.

As shown in Table 2, without extended targets, the accuracy is 65.6%, which is still higher than those using only target-independent features. After adding all extended targets, the accuracy is improved significantly to 68.2% ($p < 0.005$), which suggests that target extension does help find indi-

rectly expressed sentiments about the target. In addition, all of the three methods contribute to the overall improvement, with the head noun method contributing most. However, the other two methods do not contribute significantly.

6.3 Evaluation of Polarity Classification

Similarly, we conduct several experiments on positive and negative tweets to compare the polarity classifiers with different features, where we use 268 negative and 268 randomly selected positive tweets. The results are listed below.

Features	Accuracy (%)
Content features	78.8
+ Sentiment lexicon features	84.2
+ Target-dependent features	**85.6**
Re-implementation of (Barbosa and Feng, 2010)	83.9

Table 3. Evaluation of polarity classifiers.

From Table 3, we can see that the classifier using only the content features achieves the worst accuracy (78.8%). Sentiment lexicon features are shown to be very helpful for improving the performance. Similarly, we re-implemented the method proposed by (Barbosa and Feng, 2010) in this experiment. The results show that our system using both content features and sentiment lexicon features performs slightly better than (Barbosa and Feng, 2010). The reason may be same as that we explained above.

Again, the classifier using all features achieves the best performance. Both the classifiers with all features and with the combination of content and sentiment lexicon features are significantly better than that with only the content features ($p < 0.01$). However, the classifier with all features does not significantly outperform that using the combination of content and sentiment lexicon features. We also note that the improvement by target-dependent features here is not as large as that in subjectivity classification. Both of these indicate that target-dependent features are more useful for improving subjectivity classification than for polarity classification. This is consistent with our observation in Subsection 6.2 that most errors caused by incorrect target association are made in subjectivity classification. We also note that all numbers in Table 3 are much bigger than those in Table 1, which suggests that subjectivity classification of tweets is more difficult than polarity classification.

Similarly, we evaluated the contribution of target extension for polarity classification. According to the results, adding all extended targets improves the accuracy by about 1 point. However, the contributions from the three individual methods are not statistically significant.

6.4 Evaluation of Graph-based Optimization

As seen in Figure 1, there are several tweets which are not connected with any other tweets. For these tweets, our graph-based optimization approach will have no effect. The following table shows the percentages of the tweets in our evaluation data set which have at least one related tweet according to various relation types.

Relation type	Percentage
Published by the same person[11]	41.6
Retweet	23.0
Reply	21.0
All	**66.2**

Table 4. Percentages of tweets having at least one related tweet according to various relation types.

According to Table 4, for 66.2% of the tweets concerning the test queries, we can find at least one related tweet. That means our context-aware approach is potentially useful for most of the tweets.

To evaluate the effectiveness of our context-aware approach, we compared the systems with and without considering the context.

System	Accuracy	F1-score (%)		
		pos	neu	neg
Target-dependent sentiment classifier	66.0	57.5	70.1	66.1
+Graph-based optimization	**68.3**	**63.5**	**71.0**	**68.5**

Table 5. Effectiveness of the context-aware approach.

As shown in Table 5, the overall accuracy of the target-dependent classifiers over three classes is 66.0%. The graph-based optimization improves the performance by over 2 points ($p < 0.005$), which clearly shows that the context information is very

[11] We limit the time frame from one week before to one week after the post time of the current tweet.

useful for classifying the sentiments of tweets. From the detailed improvement for each sentiment class, we find that the context-aware approach is especially helpful for positive and negative classes.

Relation type	Accuracy (%)
Published by the same person	**67.8**
Retweet	66.0
Reply	67.0

Table 6. Contribution comparison between relations.

We further compared the three types of relations for context-aware sentiment classification; the results are reported in Table 6. Clearly, being published by the same person is the most useful relation for sentiment classification, which is consistent with the percentage distribution of the tweets over relation types; using retweet only does not help. One possible reason for this is that the retweets and their original tweets are nearly the same, so it is very likely that they have already got the same labels in previous classifications.

7 Conclusions and Future Work

Twitter sentiment analysis has attracted much attention recently. In this paper, we address target-dependent sentiment classification of tweets. Different from previous work using target-independent classification, we propose to incorporate syntactic features to distinguish texts used for expressing sentiments towards different targets in a tweet. According to the experimental results, the classifiers incorporating target-dependent features significantly outperform the previous target-independent classifiers.

In addition, different from previous work using only information on the current tweet for sentiment classification, we propose to take the related tweets of the current tweet into consideration by utilizing graph-based optimization. According to the experimental results, the graph-based optimization significantly improves the performance.

As mentioned in Section 4.1, in future we would like to explore the relations between a target and any of its extended targets. We are also interested in exploring relations between Twitter accounts for classifying the sentiments of the tweets published by them.

Acknowledgments

We would like to thank Matt Callcut for refining the language of this paper, and thank Yuki Arase and the anonymous reviewers for many valuable comments and helpful suggestions. We would also thank Furu Wei and Xiaolong Wang for their help with some of the experiments and the preparation of the camera-ready version of the paper.

References

Ralitsa Angelova, Gerhard Weikum. 2006. Graph-based text classification: learn from your neighbors. SIGIR 2006: 485-492

Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling 2010.

Christopher Burges. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2):121-167.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan S. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In Proc. of the 2005 Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005). pp. 355-362

Dmitry Davidiv, Oren Tsur and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Coling 2010.

Xiaowen Ding and Bing Liu. 2007. The Utility of Linguistic Rules in Opinion Mining. SIGIR-2007 (poster paper), 23-27 July 2007, Amsterdam.

Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision.

Vasileios Hatzivassiloglou and Kathleen.R. McKeown. 2002. Predicting the semantic orientation of adjectives. In Proceedings of the 35th ACL and the 8th Conference of the European Chapter of the ACL.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, Aug 22-25, 2004.

Thorsten Joachims. Making Large-scale Support Vector Machine Learning Practical. In B. SchÄolkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in kernel methods: support vector learning, pages 169-184. MIT Press, Cambridge, MA, USA, 1999.

Soo-Min Kim and Eduard Hovy 2006. Extracting opinions, opinion holders, and topics expressed in online news media text, In Proc. of ACL Workshop on Sentiment and Subjectivity in Text, pp.1-8, Sydney, Australia.

Ryan McDonald, F. Pereira, K. Ribarov, and J. Haji˘c. 2005. Non-projective dependency parsing using spanning tree algorithms. In Proc. HLT/EMNLP.

Tetsuya Nasukawa, Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In Proceedings of K-CAP.

Bo Pang, Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of ACL 2004.

Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques.

Ravi Parikh and Matin Movassate. 2009. Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques.

Wee. M. Soon, Hwee. T. Ng, and Danial. C. Y. Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521–544.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In proceedings of ACL 2002.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In Proceedings of AAAI-2000.

Theresa Wilson, Janyce Wiebe, Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of NAACL 2005.