IADIS International Journal on WWW/Internet
Vol. 14, No. 1, pp. 15-27
ISSN: 1645-7641

# SENTIMENT ESTIMATION OF TWEETS BY LEARNING SOCIAL BOOKMARK DATA

Yasuyuki Okamura, Takayuki Yumoto, Manabu Nii and Naotake Kamiura
*Graduate School of Engineering, University of Hyogo, 2167 Shosha, Himeji, Hyogo, 671-2201, Japan*

**ABSTRACT**

People are posting huge amounts of varied information on the Web as the popularity of social media continues to increase. The sentiment of a tweet posted on Twitter can reveal valuable information on the reputation of various targets both on the Web and in the real world. We propose a method to classify tweet sentiments by machine learning. In most cases, machine learning requires a significant amount of manually labeled data. Our method is different in that we use social bookmark data as training data for classifying tweets with URLs. In social bookmarks, comments are written using casual expressions, similar to tweets. Since tags in social bookmarks partly represent sentiment, they can be used as supervisory signals for learning. The proposed method moves beyond the basic "positive"/"negative" classification to classify impressions as "interesting", "funny", "negative", and "other".

**KEYWORDS**

Twitter, social bookmark, sentiment, machine learning, support vector machine

## 1. INTRODUCTION

People are posting huge amounts of varied information on the Web as the popularity of social media continues to increase. Twitter is a good example, where we can see tweets introducing Web pages listed in the Twitter timeline. If we want to know the reputation of the pages, however, we need to read many tweets related to the same topic, which is inefficient.

In this work, we propose a method to classify the sentiment of tweets by using machine learning. In most cases, machine learning requires a significant amount of manually labeled data. Our method is different in that we use social bookmark data as training data for classifying tweets. Social Bookmark (SBM) is a service for sharing bookmarks on the Web. In most SBM services, comments and tags are available. Users can comment on the pages they have bookmarked. In social bookmarks, comments are usually written in casual expressions, similar to tweets. Users also sometimes attach tags to the bookmarked pages to indicate the topic of the page, the user's impression, personal labels, and so on (Golder and Hubermann

2004). Tags representing impressions on the pages can be used as supervisory signals for learning the sentiment of tweets. Tweet sentiments are often modeled into the three classes of "positive", "negative", and "neutral" (Go et al. 2009). In this work, we further classify positive impressions into "interesting" and "funny".

# 2. BACKGROUND

## 2.1 Social Bookmark

SBM is a service to preserve bookmark information on the Web. In SBM, tags are used to label pages and users can comment on the pages. A single SBM can be described as $b = (u, r, t, c)$, where $u$ is a user, $r$ is a URL (resource), $t$ is a tag, and $c$ is a comment.

There has been much research related to SBM. For example, Golder and Hubermann studied the usage of tags on SBMs and classified them into seven functions (Golder and Hubermann 2004):

- identifying what (or who) it is about: topic or category
- identifying what it is: content type
- identifying who owns it: author
- refining categories
- identifying qualities or characteristics: impression or opinion
- self reference
- task organizing: e.g., to do, to read

Sen et al. (2009) and Niwa et al. (2006) focused on the first role and developed information recommendation algorithms. Yanbe et al. utilized SBM data for searching the Web (Yanbe et al. 2007), focusing on the number of SBMs on a given page and using it to define popularity. However, they did not consider the impression of users. In this work, we focus on tags expressing impressions or the characteristics of described pages (the fifth function). We use these tags for labeling the sentiment of comments automatically.

## 2.2 Related Work

Go et al. proposed a method to classify the sentiment of tweets by machine learning (Go et al. 2009), where tweets are classified as either positive or negative. In contrast, we try to classify them into more minute categories, e.g., positive tweets can be further classified into "interesting" and "funny", with tweets in the first class mainly of interest when users search for knowledge and tweets in the second class more for entertainment purposes.

One of the main contributions of our research is a new scheme for analyzing tweets by SBM data. Saito et al. tackled the same theme in a different way (Saito et al. 2012): they built a thesaurus considering the co-occurrence of SBM tags and then used it in conjunction with the feature vectors of Twitter users' profiles for user recommendation. In contrast, we propose a new learning scheme from SBM to Twitter.

Bollen et al. extracted six dimensions of mood from tweets and analyzed them with relation to economic trends (Bollen et al. 2009). They used the Profile of Mood States, which is a psychometric instrument.

Various methods to collect training data for sentiment classification automatically have been proposed. Pak and Paroubek focused on emoticons and Twitter accounts (Pak and Paroubek 2010), regarding tweets with positive emoticons as data for positive tweets and tweets with negative emoticons as data for negative tweets. They collected neutral tweets from tweets by popular newspapers and magazines. Kouloumpis et al. also used hashtags to obtain training data (Kouloumpis et al. 2011). In contrast, we use SBM data as training data for tweets.

# 3. ESTIMATING SENTIMENT ON TWITTER

To express the sentiment of a tweet relating to a Web page, we first prepare three classes: "positive", "negative", and "other". The positive class can be further classified into two subclasses:

- interesting: positive impression that users feel when they find the information is interesting
- funny: positive impression that users feel when they browse comedy or entertainment Web pages

Therefore, we ultimately classify tweets into four classes, the relationships between which are shown in Figure 1.

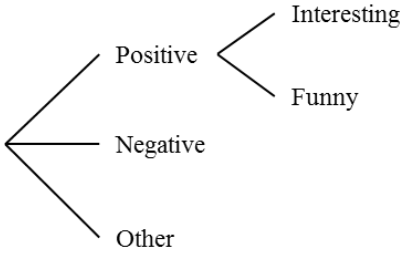


Figure 1. Sentiment classes

We estimate the sentiment of tweets in two steps:

1. Build classifiers for each class using SBM-derived data.
2. Estimate sentiment of tweets using the classifiers.

## 3.1 Building Sentiment Classifiers

### 3.1.1 Automatic Sentiment labeling by Social Tags

Instead of manual labeling on comments, we use tags in SBM, called social tags. Social tags are utilized in various ways (Golder and Hubermann 2004). Here, we focus on tags related to sentiment. We show the tags used for labeling (originally in Japanese) in Table 1. If a comment is posted with any of the tags shown in the table, this comment is automatically labeled as the corresponding class. When we make the training data, we extract only one comment per URL to avoid much effect from having a few pages with many comments.

Table 1. Tags used for automatic labeling

| Class | Tags used as positive example | Tags used as negative example |
|---|---|---|
| Positive | great, useful, joke | bad |
| Negative | bad | great, useful |
| Interesting | great, useful | joke, bad |
| Funny | joke | useful, bad |

### 3.1.2 Making Feature Vectors of Comments

We make a feature vector from a comment in SBM. Nouns and adjectives are extracted from a comment by morphological analysis. Each element corresponds to each word and its value is binary. If a certain word appears in a comment, the value of the corresponding element is 1. If the word does not appear, the value is 0. Even if the word appears several times, its value is still 1.

We use the Japanese morphological analyzer, Juman [1], which utilizes dictionaries containing emoticons. We use the appearance of emoticons as one of the features and do not differentiate between emoticon types. Therefore, the value of the feature is 1 when any type of emoticon appears and is 0 otherwise.

Comments in SBM and tweets are written using casual expressions. We consider this by taking two approaches. The first approach is to use the existence of Internet slang as one of the features. We focus on the repeating "w" expression, which is Japanese Internet slang that roughly translates as the English "lol". For example, "ww" and "wwww" have almost the same meaning. Therefore, we adopt the existence of repeating "w" as a feature. If repeating "w" appears in a comment, the value of the element is 1, and otherwise, it is 0. Exceptions are the "www" that occurs in a URL pattern such as http://www.domain.com/.

In the second approach, we consider any mistakes in morphological analysis. In Japanese sentences, words aren't separated by spaces, so we need to tokenize sentences to extract words. Popular Japanese morphological analyzers are trained with sentences in formal expressions such as those found in newspaper articles. Such analyzers are not able to tokenize sentences written in casual expression, which might have a negative effect on the classification. Here, we use the Japanese morphological analyzer Mecab (Kudo et al. 2004) and compare its results with those by Juman. If there is obvious disagreement between the results of tokenization, we do not use the words. If a tokenization disagreement occurs in a sentence, it might indicate that the sentence is written in a casual expression, which is a clue for estimating sentiment. Therefore, we also use the existence of disagreement between the morphological analyzers as one of the features. If there is disagreement, the value of the feature is 1, and otherwise, it is 0. We denote this method as "disagreement".

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

## 3.2 Estimating Sentiment of Tweets

### 3.2.1 Making Feature Vectors of Tweets

A feature vector of a tweet is made the same way as the feature vectors of comments in SBM. In preprocessing, we remove URLs and Web page titles from the tweets. The title patterns are as follows.

- Text surrounded by the title tags in the linking Web page.
- Text in a data-text attribute. Twitter Inc. recommends this style to embed the title of the Web page in a tweet[2].
- Specific parameters in the linking URL such as blogs, news sites, and shopping sites

Unfortunately, these patterns do not cover all potential title patterns. It is particularly difficult to identify the title part in tweets when users summarize the title. Addressing this remains a future work.

### 3.2.2 Sentiment Estimation by SVM

We use a support vector machine (Vapnik 1998) to construct a classifier for positive ($M_{pos}$), a classifier for negative ($M_{neg}$), and a classifier for "funny"($M_{fun}$). We combine these three classifiers to classify the sentiment of tweets. First, we decide whether a given tweet is positive or negative by $M_{pos}$ and $M_{neg}$. If the tweet is positive, we further classify it into "interesting" or "funny". This algorithm is shown in Table 2. Each $\theta_{class}$ is a threshold to decide each $class \in \{pos, neg, fun\}$. $D$ is a feature vector of a tweet and $C$ is the estimated class of $D$. $\mathrm{predict}(M_{class}, D)$ means the probability that $D$ is predicted as *class* by $M_{class}$. The probability is calculated by Wu et al.'s method (Wu et al. 2004).

## 4. EXPERIMENTS

For the evaluation, we use F-measure defined by precision and recall. Precision, which is the ratio of correct estimations among the estimated results, is defined as

$$Precision = \frac{R_{c \to c}}{R_{* \to c}}$$

,

where $c \in \{\mathrm{positive}, \mathrm{negative}, \mathrm{funny}, \mathrm{interesting}\}$. $R_{c \to c}$ means the number of tweets whose actual class is class *c* and that are classified as *c*. $R_{* \to c}$ means the number of the tweets that are classified as class *c*. Here, we regard the "positive" class as a superclass of the "interesting" and "funny" classes. Recall is the ratio of correct estimations among the answer data. We show the definition in

$$Recall = \frac{R_{c \to c}}{R_{c \to *}}$$

[2] http://dev.twitter.com/docs/tweet-button

Table 2. Algorithm for estimating sentiment

**Input** : $M_{pos}$, $M_{neg}$, $M_{fun}$, $D$, $\theta_{pos}$, $\theta_{neg}$, $\theta_{fun}$
**Output** : $C$
$P_{pos} \leftarrow \text{predict}(M_{pos}, D)$
$P_{neg} \leftarrow \text{predict}(M_{neg}, D)$
**if** $P_{pos} > \theta_{pos}$ AND $P_{neg} < \theta_{neg}$ **then**
  $P_{fun} \leftarrow \text{predict}(M_{fun}, D)$
  **if** $P_{fun} > \theta_{fun}$ **then**
    $C \leftarrow funny$
  **else**
    $C \leftarrow interesting$
  **end if**
**else if** $P_{neg} > \theta_{neg}$ AND $P_{pos} < \theta_{pos}$ **then**
  $C \leftarrow negative$
**else**
  $C \leftarrow other$
**end if**
**return** $C$

where $R_{c\rightarrow *}$ means the number of tweets whose actual class is class *c*. F-measure is the harmonic mean of precision and recall, defined as

$$F = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

We used LIBSVM (Chang and Lin 2011) and adopted the RBF kernel. We set the parameters $\theta_{pos} = \theta_{neg} = \theta_{fun} = 0.5$.

## 4.1 Dataset

We prepared SBM data and tweet data. SBM data was obtained from Hatena bookmark, a popular SBM service in Japan[3]. The details of the obtained data are provided in Table 3. We randomly selected 1200 examples for each class in Table 3 for a total of 4800 examples. Several comments in the SBM data were then manually given with sentiment labels, shown in Table 4.

Table 3. Collected SBM data

| Type | No. |
|---|---|
| Unique URLs | 18,687 |
| URLs with comments | 15,876 |
| Comments | 340,595 |

[3] http://b.hatena.ne.jp/

Table 4. Labeled SBM data

| Class | Number of comments |
|---|---|
| Negative | 741 |
| Funny | 545 |
| Interesting | 1,409 |

We collected tweets by using the Twitter Streaming API[4]. First, we collected 5.4 million tweets in Japanese and narrowed them down to tweets mentioning Web pages and that had been posted by humans. We eliminated automatically posted tweets in several ways. First, retweets and replies were excluded. To eliminate tweets linked to prize promotion sites, we removed tweets containing words such as "gift" and "invite." We also used only Twitter clients that are used to post tweets, thus removing bots. Ultimately we were left with 53,884 tweets.

We randomly selected 5000 tweets, and four human labelers labeled them with sentiment labels. Each tweet was labeled by two labelers. If the results are the same, the tweet with the label was adopted in the dataset. We show the detail of the dataset in Table 5. The unadopted tweets and the reasons are shown in Table 6, and the tweets with different labels are shown in Table 7.

Table 5. Labeled tweets

| Class | No. |
|---|---|
| Negative | 523 |
| Funny | 407 |
| Interesting | 389 |
| Total | 1,319 |

Table 6. Unadopted tweets

| Reason | No. |
|---|---|
| Advertisements or spam | 448 |
| No sentiment | 1,355 |
| Disagreement between labelers | 1,878 |
| Total | 3,681 |

Table 7. Tweets with different labels

| Labels | | No. |
|---|---|---|
| Negative | Funny | 49 |
| Funny | Interesting | 167 |
| Interesting | Negative | 74 |
| Negative or Funny or Interesting | Ads or spam or no sentiment | 1,588 |

[4] https://dev.twitter.com/docs/api/streaming

## 4.2 Evaluation for Estimating Tweets

To analyze the effectiveness of the features, we compared the F-measures of the case when we used the different features. The results are shown in Figure 2. The results with all features are best. The features derived using two different morphological analyzers had a positive effect on all classifiers. The "Internet slang" feature also improved all the classifiers.

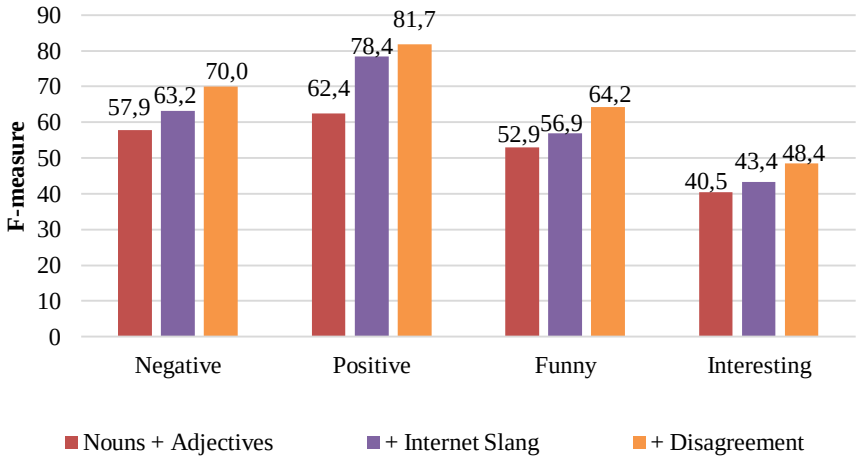


Figure 2. Results for estimating sentiment of tweets

In contrast, the F-measure of the "interesting" class is lower than the ones of the other classes. We show the precision and the recall in Table 8. From the results, the precision of the "interesting" class is especially low. We investigated what types of errors were frequent. The results are shown in Table 9. The 42 tweets classified as "other" are not in the table. The most frequent errors are that "funny" tweets are classified as "interesting". It causes the low precision of the "interesting" class. However, the F-measure of the "positive" class is high. This means the "funny" classifier should be improved.

Table 8. Precision, recall and F-measure when we used all features

| | Negative | Positive | Funny | Interesting |
|---|---|---|---|---|
| Precision | 66.27% | 84.62% | 73.21% | 44.80% |
| Recall | 74.12% | 79.05% | 57.17% | 52.66% |
| F-measure | 69.98% | 81.74% | 64.21% | 48.41% |

Table 9. Confusion matrix for sentiment estimation

| | | Answer | | |
|---|---|---|---|---|
| | | Interesting | Funny | Negative |
| Classifier | Interesting | 168 | 128 | 79 |
| | Funny | 66 | 287 | 39 |
| | Negative | 85 | 87 | 338 |

## 4.3 Comparison with SBM Comment Classifier

To analyze the effect of using SBM data for tweet classification, we classified SBM comments by the proposed classifiers. The data in Table 4 are used as test data and the F-measures (%) are shown in Figure 3. The results are best when all features are used. This tendency is also observed when tweets are classified. Although all of the results are better than the results of the proposed tweet classifications, the differences are small in the "positive", "negative" and "funny" classes. However, in the "interesting" class, the proposed method is much worse than the SBM comment classifier. This may be caused by the difference of the tendency between Twitter and Hatena bookmarks. We plan to compare the usage of these services as future work.

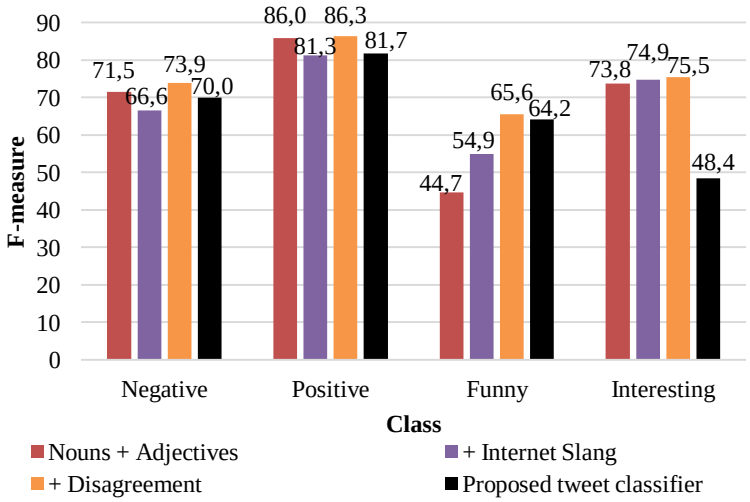


Figure 3. Results for estimating sentiment of SBM comments

## 4.4 Comparison with Tweet Classifier Learned by Tweets

In order to analyze the effect of using SBM data for tweet classification from another side, we compare the proposed method with classifiers learned by tweets. First, we built the following three classifiers using tweets as training data.

- The classifier for the positive class: "interesting" and "funny" tweets were used as positive examples and "negative" tweets were used as negative examples.

- The classifier for the negative class: "negative" tweets were used as positive examples and "interesting" and "funny" tweets were used as negative examples.
- The classifier for the funny class: "funny" tweets were used as positive examples and "interesting" and "negative" tweets were used as negative examples.

These classifiers were combined in the same way as the proposed method (see Table 2), and the sentiment classifier learned by tweets was built.

We used 385 tweets for each class and conducted 10-fold cross validation. The results are shown in Figure 4. From the results, the proposed method marks comparable F-measure with the classifiers learned by tweets in the "positive", "negative" and "funny" classes. Next, we changed the size of the training data (100, 200, 300, All) and compared the F-measures. The results are shown in Figure 5. When the size of the training data is less than 300, the proposed methods are better in the "positive", "negative" and "funny" classes, and it is comparable in the "interesting" class with the classifiers learned by tweets. If large number of labeled tweets are available, it is better to use the labeled tweets as training data. However, in our experiment, the tweets with sentiment was only 25% out of 5000 tweets (see Table 5). This means that we have to label quadruple number of tweets that are enough for training data. In contrast, our method needs no labeled data, and it has performance comparable to the classifiers learned by labeled data. Therefore, our method is practical to build a sentiment classifier.

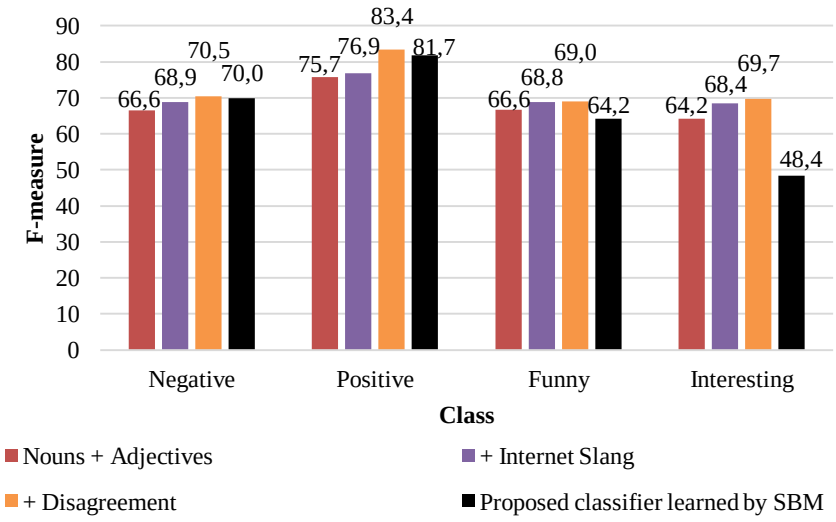


Figure 4. Results for estimating sentiment of tweets learned by SBM with different features

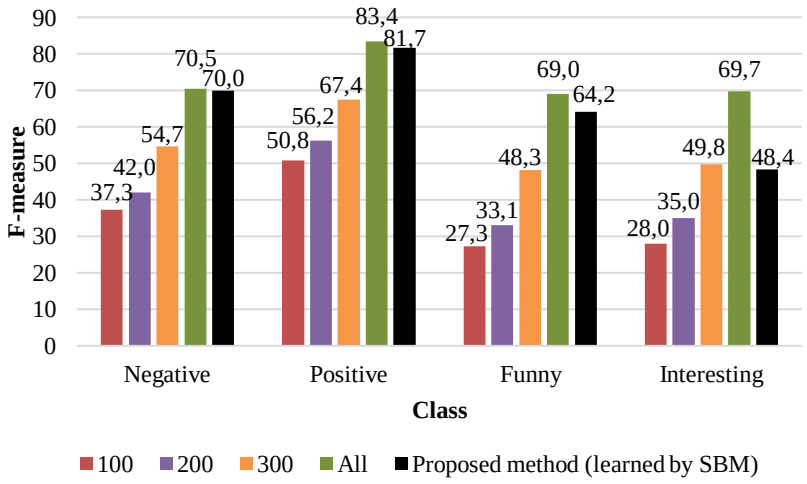


Figure 5. Results for estimating sentiment of tweets learned by SBM with different size of training data

## 4.5 Parameter Optimization

In the above experiments, we used the parameters $\theta_{pos} = \theta_{neg} = \theta_{fun} = 0.5$. To improve the performance, we investigated the optimal parameters in the following steps:

- We respectively changed $\theta_{pos}$ and $\theta_{neg}$ from 0.100 to 0.900 and we adopted the values where the harmonic mean of F-measures of the "positive" class and the "negative" class was best.
- We changed $\theta_{fun}$ from 0.100 to 0.900 and we adopted the value where the harmonic mean of F-measures of the "interesting" class and the "funny" class was best.

The optimal parameters were $\theta_{pos} = 0.446, , \theta_{neg} = 0.543$ and $\theta_{fun} = 0.754$. The results are shown in Table 10 and the comparative results with the default parameters are shown in Figure 6. We show the distribution of each class in Table 11. The F-measures of all classes were improved. Especially, the lower $\theta_{pos}$ improved the recall of the "positive" class, and the higher $\theta_{neg}$ improved the precision of the "negative" class. Although the higher $\theta_{fun}$ improved the recall of the "interesting" class, the precision was still low. One of the reasons is that short tweets tend to be classified as "funny".

Table 10. Precision, recall and F-measure when we used optimized parameters

| | Negative | Positive | Funny | Interesting |
|---|---|---|---|---|
| Precision | 72.31% | 83.70% | 75.85% | 46.94% |
| Recall | 70.22% | 85.04% | 57.80% | 66.99% |
| F-measure | 71.25% | 84.37% | 65.61% | 55.20% |

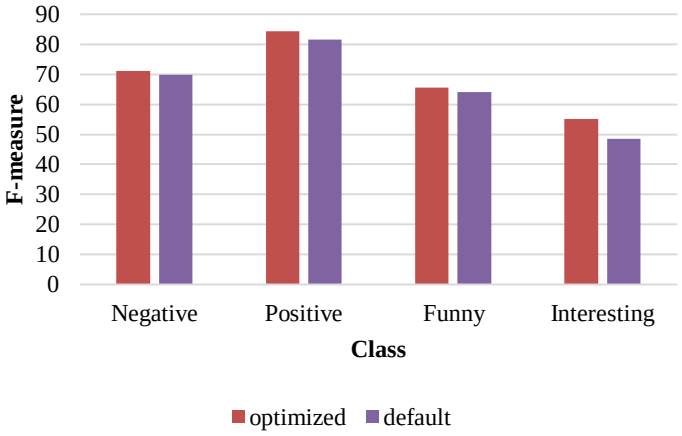


Figure 6. Results with optimized parameters and default parameters

Table 11. Confusion matrix for sentiment estimation when we used optimized parameters

| | | Answer | | |
|---|---|---|---|---|
| | | Interesting | Funny | Negative |
| Classifier | Interesting | 207 | 144 | 90 |
| | Funny | 48 | 285 | 44 |
| | Negative | 54 | 67 | 316 |

## 5. CONCLUSION

We proposed a method to classify tweets into four types of impression: “interesting”, “funny”, “negative”, and “other”. We built tweet classifiers for each sentiment by automatically obtaining training data from SBM. In this data, comments are converted into feature vectors, and if specific tags are used, the comments are labeled as the corresponding sentiment. The experimental results showed that SBM data can be utilized as training data for classifying the sentiment of tweets. Although it marked slightly worse F-measure than the classifiers learned by labeled tweets, it is available without laborious labeling tasks. As future work, we will apply our method in estimating reputation of Web pages.

## ACKNOWLEDGEMENT

This work was supported by Grant-in-Aid for Young Scientists (B) Grant Number 24700097 from Japan Society for the Promotion of Science.

## REFERENCES

Bollen, J. et al., 2009, Modeling Public Mood and Emotion: Twitter Sentiment and Socio-economic Phenomena. *Computing Research Repository*, abs/0911.1583.

Chang, C.C. and Lin, C.J, 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27.

Go, A. et al. 2006. Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report*, Stanford, pp. 1–12.

Golder, S.A. and Huberman, B.A. 2006. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science* 32(2), pp.198–208.

Kouloumpis, E. et al., 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, pp. 538–541.

Kudo, T. et al., 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, pp. 230–237.

Niwa, S. et al., 2006. Web Page Recommender System based on Folksonomy Mining for ITNG' 06 submissions. *Proceedings of the Third International Conference on Information Technology*, Las Vegas, USA, pp. 388–393.

Pak, A., Paroubek, P., 2010 Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, pp. 1320–1326.

Saito, J. and Yukawa, T., 2012. Extracting User's Interest based on Social Bookmark Tags. *Advances in Smart Systems Research* 1(1), pp.7–12.

Sen, S. et al., 2009. Tagommenders: Connecting Users to Items through Tags. *Proceedings of the 18th International Conference on World Wide Web*. Madrid, Spain, pp.671–680.

Vapnik, N. V., 1998. *Statistical Learning Theory*. Wiley-Interscience, New York, USA.

Wu, T.F. et al., 2004. Probability estimates for Multi-class Classification by Pairwise Coupling. *The Journal of Machine Learning Research*, Vol.5, pp.975–1005.

Yanbe, Y. et al., 2007. Can Social Bookmarking Enhance Search in the Web? *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. Vancouver, Canada, pp. 107–116.