This article should be cited as:
Yu Suzuki, Filtering Method for Twitter Streaming Data using Human-in-the-Loop Machine Learning, Journal of Information Processing Vol.27(2019) (online)

IPSJ Transactions on Databases Vol.12 No.4 1–8 (Apr. 2019)

Regular Paper

Filtering Method for Twitter Streaming Data using Human-in-the-Loop Machine Learning

Yu Suzuki[1,a)]

Received: xx xx, xxxx, Accepted: xx xx, xxxx

Abstract: A large number of texts is posted daily on social media. However, only a small portion of these texts is informative for a specific purpose. For example, in order to collect a set of tweets for marketing strategy, we should collect a large number of tweets related to a specific topic with high accuracy. If we accurately filter the texts, we can continuously obtain fresh and useful information in real time. In a keyword-based approach, filters are constructed using keywords, but selecting the appropriate keywords is often tricky. In this work, we propose a method for filtering texts that are related to specific topics using a classification method that is based on crowdsourcing and machine learning. In our approach, we construct a text classifier using fastText and then annotate whether the tweets are related to the topics using crowdsourcing. For constructing an accurate classifier, we should prepare a large amount of learning data. However, this process is costly and time-consuming. To construct an accurate classifier using a small number of learning data, we consider two strategies for selecting tweets which the crowdsourcing participants should assess: optimistic and pessimistic approach. Then, we reconstruct the text classifier using the annotated texts and classify them again. If we continue instigating this loop, the accuracy of the classifier will improve, and we will obtain useful information without having to specify the keywords. Experimental results demonstrate that our proposed system is adequate for filtering social media streams. Moreover, we discovered that the pessimistic approach is better than the optimistic approach.

Keywords: Information filtering, Human-in-the-loop, Human factor, Machine learning

1. Introduction

A massive number of texts is posted on social network services, and information about real situations can be gleaned from these texts. There have been many studies on how to extract the necessary information from these data [2]. Information filtering [3] is a process of retrieving texts from text streaming, and the keyword-based method is often used for this. However, it is difficult to choose the appropriate keywords that correspond to the information needed for two reasons: 1) information needed is generally vague and 2) sentences on social networks are often grammatically broken. For example, if a user collects tweets about Kyoto sightseeing, the user should specify the keywords as not only "Kyoto" and "Sightseeing" but also "temple" and "Kinkaku-Ji," the name of a places in Kyoto. However, we should make complex queries using not only AND operator but also OR and NOT, in order to filter tweets about the temples which are not in Kyoto.

In our research, we have developed systems that collect a small number of texts relevant to subjective queries from a large text stream by using machine learning and crowdsourcing. Our objective is to extract from text streaming the tweets related to a specific topic. The topics we assume are general, and the extracted tweets are relevant for many users (i.e., not personalized topics).

Machine learning is one method for predicting the label of unknown data by generalization with training data as an input. Classification accuracy will be low if the quality of the input is low [4, 5]. The monetary and processing cost of machine learning is low. Therefore, to improve the classification accuracy of the classifier, it is necessary to prepare a huge amount of good quality data as a training data.

Crowdsourcing is useful in processing tasks that require human intuition, such as sentiment analysis [6] and translation [7]. Although crowdsourcing enables the collection of high-quality data, it is expensive and time-consuming compared to machine learning. Assessing all tweet streaming data by crowdsourcing is unrealistic because there is a large number of tweets and because paying the crowdsourcing workers for assessing all tweets would be very costly.

To accomplish the goal, we combine two techniques: crowdsourcing and machine learning, which is called a human-in-the-loop model [8]. By using human-in-the-loop, we compensate for the disadvantages of both machine learning and crowdsourcing. In this framework, crowdsourcing is used for creating the training data to improve the classifier and cleaning the output of the system, and machine learning is used for reducing the number of tweets to be judged by crowdsourcing.

In general, almost all tweets in a twitter stream are irrelevant to a specific topic. In our experiment, we set the topic as "Kyoto sightseeing." In this case, the number of relevant tweets is less than 0.1%. However, people assessing the tweets are hu-

[1] Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[a)] ysuzuki@is.naist.jp
[*1] This paper is an extended version of [1].

man beings, so if they think that many tweets are irrelevant to the topic, they may apply *irrelevant* to all the tweets without browsing. As a result, we will miss the relevant tweets even if we use expensive crowdsourcing mechanism. Therefore, we have to investigate whether people can make appropriate judgment of these heavily biased data. The results of experiment 1 described in section 4.1 confirm that crowdsourcing workers can appropriately judge irrelevant tweets.

In machine learning, the quality of the outputs depends on the quality of the learning data, which suggests that the more tweets are being presented to people, the more accurately they can obtain the necessary tweets. However, preparing a large amount of learning data is costly and time-consuming. Therefore, in order to create an accurate classifier using a small number of learning data, it is important to decide which tweets should be assessed. In this paper, we propose two approaches for selecting unlabeled tweets: *optimistic* and *pessimistic*. In the optimistic approach, people assess the tweets as relevant and place them near the decision boundary of the classifier. In the pessimistic approach, people assess the tweets as relevant and place them far from the decision boundary of the classifier. Optimistic approach is widely used in the systems based on active learning. However, when we observe the behavior of the dataset, we discover that many tweets which are classified as relevant by the classifier are not in fact relevant. Therefore, to converge the models quickly, we should select the tweets with high relevance possibilities.

Contributions of this paper are as follows.

- By combining machine learning and crowdsourcing, it became possible to filter relevant tweets without using keywords.
- By presenting relevant tweets using pessimistic approach, the collection of target tweets became more efficient.

2. Related Work

In social network services, numerous useful texts are posted, such as those related to the spread of influenza and the occurrence of an accident. Several systems have been proposed to capture these incidents quickly [9] and thereby to utilize social media services as a kind of social sensor. Many information-filtering methods for gleaning useful information from streaming data have been proposed, and many of them are used in systems generated for personalization [10], in what is called "personalized information filtering." These techniques are based on information retrieval and are not appropriate for short texts such as tweets. Therefore, bag-of-words features along with domain-specific knowledge [11], the relationship between users [12], and user behaviors such as re-tweeting [13] are used as features to filter tweets.

Distributed expressions using words with vectors contribute greatly to this development. word2vec [14–16] and fastText [17] and the like construct distributed expressions using neural networks. A method of document classification using such distributed expressions has been proposed [18]. However, as far as the authors know, there is no way to achieve information filtering using document classification by distributed representation. Therefore, we propose an information filtering method that combines crowdsourcing and document classification.

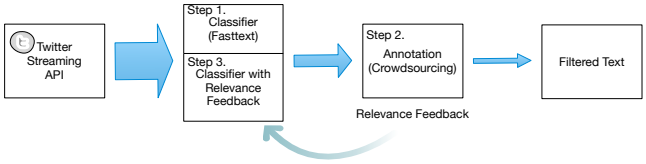


Fig. 1 Overview of proposed system.

Relevance feedback is important for improving the accuracy of information filtering. Rocchio [19] proposed a relevance feedback mechanism on the vector space model. In this mechanism, accuracy improves as more feedback is given, so a method using crowdsourcing for feedback has been proposed [20, 21]. However, in this method, it is assumed that the set of documents that are compatible is sufficiently large compared with the whole set, so it was not clear whether it can be used for information filtering. In our research, we clarify whether relevance feedback by crowdsourcing is effective when the relevant documents are extremely limited.

The critical issue of the goal is to find relevant tweets using a small number of learning data. Jörger et al. [22] proposed a method of combining multiple strategies for collecting training data with ε-greedy algorithm. However, before using this method, we should first understand the features of the outputs using each strategy. One contribution of this paper relative to Jörger's work is that we discovered the outputs using each of the two strategies (optimistic and pessimistic.)

Ertiken et al. [23] and Zhang et al. [24] tackled the problem of using imbalanced data as the input, which is similar to the problem we faced. However, the common goal of these methods is to generate accurate classifiers after the relevance feedback mechanism is converged. We suppose that it takes many steps to converge the classifier, and in our experiment, the feedback mechanism is not converged. Therefore, we treat this issue as future work.

3. Information Filtering Method

Our proposed information filtering system uses a combination of crowdsourcing and machine learning techniques. An overview is shown in Fig. 1.

(1) Build a classifier using labeled data. (Sec. 3.1)
(2) Classify tweets from tweet streams and obtain relevant tweets. (Sec. 3.2)
(3) Select tweets which should be annotated by the crowdsourcing workers (Sec. 3.3) and go to (1).

We explain each step in the following section.

3.1 Build Classifier

We use fastText [17, 25], an application for word embedding and classification, for classifying tweets. The inputs of fastText for building a classifier are a training set, a pre-trained embedding model, and parameters. First, we prepare a dataset for constructing the initial classifier. We remove tweets from the dataset if they include URLs or mention other users. Then, we prepare a set of tweets $T = \{t_1, t_2, \cdots, t_N\}$ that are related to specific topics using crowdsourcing. Each tweet is given a label *relevant* or *irrelevant*.

Next, we extract bag-of-word features from the texts. We use MeCab [*1] with the IPADIC-Neologd dictionary [*2] as morphological analysis to extract words. We assume that there are many new words in the tweets, so we use a dictionary called neologd that includes new words. Then, we extract nouns, verbs, adjectives, and adverbs. The feature vectors $f(t)$ of tweet t are defined as follows:

$$f(t) = [d_1(t), d_2(t), \cdots, d_\ell(t)] \subset \mathbb{R}^{\mathcal{D}} \quad (1)$$

where $\mathcal{D}$ is the dictionary size and $d_i(t)$ $(i = 1, 2, \cdots, \ell)$ is the frequency of occurrence of the i-th term. We use a pre-trained distributed representation [*3] constructed using the Japanese Wikipedia corpus [*4] for generating the classifier. Using this distributed representation data, we can consider the synonyms.

Finally, using the classifier included in fastText implementation [26], we construct the classifier of the tweets.

3.2 Classify Tweets

Next, we classify the tweets obtained from text streaming as *relevant* or *irrelevant* by using the classifier described in Section 3.1. The label *relevant* means that the tweets are relevant to the topic, and the label *irrelevant* means that the tweets are irrelevant to the topic. At this time, the classifier also calculates the certainty level $c(t)(0.5 \leq c(t) \leq 1)$ of the label for the tweet t. For example, if a tweet t is labeled as "relevant" and $c(t)$ equals to 0.7, t is relevant with 70% possibility. Therefore, if $c(t)$ is 0.5, the classifier cannot decide whether t is relevant or not. The lower bound of $c(t)$ depends on the number of categories. In this paper, we consider two categories, relevant and irrelevant, for classification. However, if we should classify tweets into k categories, the lower bound of $c(t)$ is $1/k$.

3.3 Annotation

Next, we manually confirm if the tweets classified as *relevant* / *irrelevant* by the classifier were truly relevant / irrelevant, using crowdsourcing.

As shown in Fig. 2, although accuracy can be improved by making judgments (and the more people, the better the accuracy), the accuracy of the final classifier improves as the number of judged tweets increases, even if the accuracy is low. Each judgment was considered final (i.e., there was no redundancy by multiple crowdsourcing workers or averaging that took place).

The accuracy of the classifier changes depending on the training set, this means which tweets was judged by a worker. If we prepare only irrelevant tweets to construct the classifier, the accuracy of the classifier will improve a little. Campbell et al. [27] pointed out that classifiers perform differently depending on the sample selection method. We consider the following two approaches: optimal and pessimistic.

3.3.1 Strategy 1: Optimistic Approach

When a classifier judges tweets, some tweets are difficult to judge. The certainty level $c(t)$ for these tweets by the classifier is around 0.5, the lowest value. Therefore, the idea of this strategy is that if these tweets are manually judged by crowdsourcing with high priority, the accuracy of the classifier will quickly improve. Lewis et al. [28] proposed a technique called uncertainty sampling, which Simon et al. [29] later used in an application [for? with?] support vector machines. In this method, tweets that are near the decision boundary are annotated. The blue part in Fig. 2 shows the tweets which should be annotated using this approach.

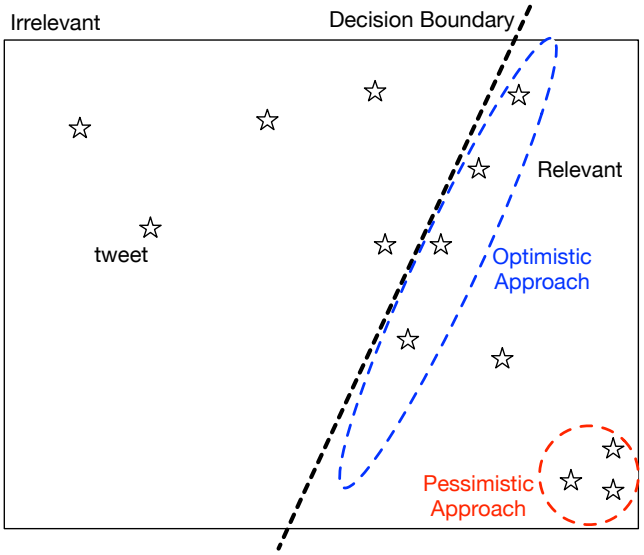


Fig. 2 Intuitive example of our proposed method.

This strategy is widely used in existing methods, for example, in active learning. It is considered effective in situations where truly relevant and irrelevant tweets are mixed in the vicinity of the decision boundary. In other words, it is best used in situations where the decision boundary during the iteration is close to the true decision boundary. Here, we call *true decision boundary* the decision boundary which can completely divide the relevant and irrelevant tweets. However, if the true decision boundary and the decision boundary during the iteration are far apart, and if the true decision boundary is close to the tweets which the classifier marks as relevant with 100% probability, the tweets which are selected by strategy 1 are not always appropriate.

To solve this issue, we propose a new strategy described in the following section.

3.3.2 Strategy 2: Pessimistic Approach

The main idea of this strategy is that we select the tweets which are labeled as relevant by the classifier with high certainty and have them assessed by the crowdsourcing workers.

In our experiment, less than 0.1% of all tweets should be considered as irrelevant. This means that even if a tweet is marked as relevant and the certainty level is high, the tweet should be judged as irrelevant with high possibility. Therefore, in this strategy, we select the related tweets with a high certainty level. The points in the red circle in Fig. 2 show the tweets which should be annotated using this approach.

This method is considered suitable for when the classification performance by classifiers is not sufficient for the unlabeled tweets. Therefore, when we use a classifier that can accurately

[*1] http://taku910.github.io/mecab/
[*2] https://github.com/neologd/mecab-ipadic-neologd
[*3] https://qiita.com/Hironsan/items/513b9f93752ecee9e670
[*4] https://dumps.wikimedia.org/jawiki/20170101/

classify this kind of unbalanced labeled tweets, this strategy is not suitable. However, to the best of our knowledge, a classifier for unbalanced labeled tweets does not exist.

3.4 Output

Only the tweets that people judged as relevant in Section 3.3 are considered as final. At the same time, tweets determined as relevant or irrelevant here were used as training data for the classifier construction in Section 3.1.

4. Evaluation

We performed two experiments. Experiment 1 is an evaluation of the participants. Since they were human beings and therefore different from machines, there was a possibility that the correct answer rate could differ in the ratio of "relevant" and "irrelevant." We performed this experiment to determine if this was indeed the case. In experiment 2, we used actual crowdsourcing with a machine-learning based classifier to compare which of the two strategies, optimistic or pessimistic, performs with better accuracy. In this experiment, we used the tweets written in Japanese.

4.1 Evaluation 1: Worker Behaviors

First, we evaluate the performance of human characteristics. In the proposed method, the participants should assess the tweets. However, almost all tweets are irrelevant to the specified topics. In this situation, we suppose that many participants may judge relevant tweets as irrelevant or judge all tweets as irrelevant. To find whether this issue exists or not, we conducted an experiment to see if this imbalance affects the accuracy rate.

4.1.1 Experimental Setup

We gathered in advance a set of tweets that would be correct. Through crowdsourcing, we collected 3,580 tweets related to sightseeing in Kyoto. In this task, participants search tweets related to Kyoto sightseeing using the official twitter Web search interface with the keywords selected by themselves.

We did not use any tweets containing images, URLs, etc. For each tweet, we made an assessment with more than ten participants; tweets that more than half of the participants judged as related to sightseeing in Kyoto were considered as correct, and the other tweets as incorrect. As a result, the number of correct tweets was 771 and the number of incorrect tweets was 2, 809. We randomly selected relevant and irrelevant tweets from this set. We prepared 21 groups with 5% increments from 0% to 100% proportion of relevant tweets. If the ratio was 50%, 50% of relevant tweets was included.

We hired 114 people using CrowdWorks [website]. Each participant was assigned an ID number. Then, we created 21 groups based on the participants' ID. Each group corresponded to the correct answer rate. When the ID number divided by 21 was m, we set the correct answer rate to $m/20$ for that person. In other words, when the ID of a person was 100 (i.e., when $m = 16$), the correct answer rate was set to $16/20 = 0.8$. Since this correct answer rate always assigned the same value for each worker, the correct answer rate did not change during the experiment.

The participants judged 50 tweets and assign a relevant or irrelevant label to each tweet. We then compared the accuracy of these labels with the correct answers that had been prepared in advance.

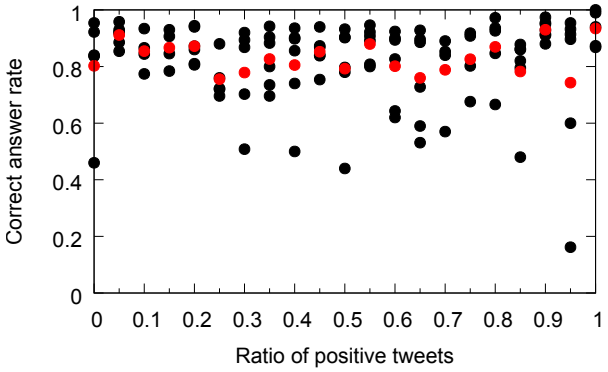


Fig. 3 Ratio of relevant tweets to correct answer rate.

4.1.2 Results

The results are shown in Fig. 3. Each black point corresponds to one participant. The average correct answer rate was 0.83, and the correct answer rate for each ratio is indicated by a red point. We found that the correct answer rate for each ratio of relevant tweets did not diverge from the average. To confirm this statistically, we performed ANOVA testing on the assumption that there is a correlation between the ratio of relevant tweets and the correct answer rate.

Using ANOVA, we found that the p-value was 0.51. In other words, we could not conclude that there was a significant difference in the correct answer rate for each ratio of relevant tweets. This indicates that even if the percentage of the correct answers was quite biased, there was no big difference in the rate at which the accurate work had been done. Therefore, in the next stage, evaluation was conducted by machine learning incorporating human judgment.

4.2 Evaluation 2: Classifier Accuracy

We used the optimistic and pessimistic approaches (Sec. 3.3) to determine the effectiveness of these strategies using the number of adequate tweets to be obtained.

4.2.1 Data

We prepared two groups of tweet data: labeled and unlabeled. For labeled data, we used all 3,580 manually collected tweets described in Sec. 4.1.1. For unlabeled data, the Twitter Streaming API was used to collect more than 1 million tweets in advance. To ensure the same conditions when comparing two strategies, we used the same tweets. Therefore, if the performance of the classifier was the same, the same tweet was extracted.

4.2.2 Procedure

Evaluation experiments were carried out as follows.

- Build a classifier using labeled data.
- Arrange unlabeled data in chronological order and classify using a classifier. Obtain 1,000 relative tweets.
- Classify tweets against relative tweets by crowdsourcing.
- Add judgment results of crowdsourcing to labeled data and return to 1.

First, we classify tweets by using the optimistic strategy and then we classify tweets again by pessimistic strategy. The num-

Table 1 Parameters for fastText.

Parameter	Value
Number of epochs	10,000
Size of vectors	300
Number of buckets	100,000,000
Loss function	Negative sampling
Number of negatives sampled	10
Minimum number of word occurrences	1
Max length of word n-gram	1
Learning rate	0.075

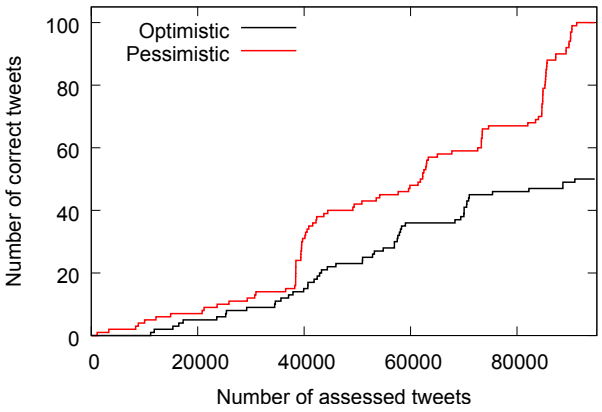


Fig. 4 Ratio of relevant tweets and correct answer rate.

Table 2 Experimental setting.

	Optimistic	Pessimistic
Manually processed tweets	94,597	176,238
No. of workers	72	72

ber of participants was the same in both strategies, but the participants were different. We confirmed that the correct ratio of each participant in both systems was almost the same. In our experiment, if the correct ratio of a participant is extremely low, we lock the account of that participant, so that he or she cannot continue doing the assessment.

4.2.3 Results and Discussion

The hyperparameter used for the classifier is shown in Table 1. In a preliminary experiment using initially labeled data, parameters that can classify [tweets as] relevant or irrelevant with high accuracy were obtained by grid search. We used the same hyperparameters for the systems using two strategies.

We obtained 94,598 and 176,238 assessments by using the optimistic and pessimistic approaches, respectively. For comparison, we used all 94,548 assessments by the optimistic approach and the first 94,548 [of 176, 238 ?] assessments by the pessimistic approach. The results are shown in Fig. 4. We discovered that the correct tweet could be collected twice as fast by the pessimistic approach than by the optimistic approach. Specifically, when the number of the assessed tweets was 40,000, the system could collect many correct tweets when it exceeded 85,000.

Figure 5 shows the number of model reconstructions (steps) vs. the ratio of relevant tweets. In this figure, a point shows the percentage of new tweets a classifier judged as relevant at each step. For example, the value at step 3 is 0.1 (red point), this means that the classifier judges the tweets as relevant at the ratio of 1 to 10 when the classifier was rebuilt three times. From this figure, in the pessimistic approach, many tweets were judged as relevant in the first three steps, but after step 4, the ratio is lower than in

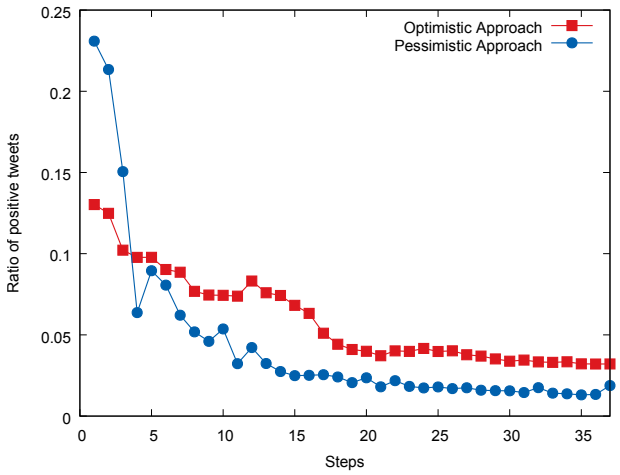


Fig. 5 The number of steps vs. ratio of relevant tweets

the optimistic approach and is converged to about 0.02. On the other hand, in the optimistic approach, in the first step, the classifier judges a smaller number of relevant tweets than that by the pessimistic approach. However, the ratio of relevant tweets does not decrease in subsequent steps. The same number of relevant tweets is selected at each level. Therefore, decreasing the ratio of relevant tweets means that it is possible to judge many tweets in a short time. As a result, it was found that the pessimistic approach can handle many more tweets compared with the optimistic approach.

From Figs. 4 and 5, the accuracy of the initial classifier is not enough for filtering texts. At step 1 in Fig. 5, about 23% of tweets are selected as relevant by the classifier of the pessimistic approach. We manually analyzed the filtered tweets and found only two relevant tweets from 10, 000 tweets. In Fig. 4, we found that there is only a small number of correct tweets when the crowdsourcing workers assess a small number of tweets. Therefore, the workers should process many tweets, and the requesters should pay more wages to the workers.

In this experiment, we were able to collect as much as 50 and 100 relevant tweets for the optimistic and pessimistic approach, respectively, for 4,500 JPY (45 USD). At that time, we got an assessment result of around 95,000 tweets. Therefore, we confirmed that we could find relevant tweets at low cost.

Examples of actual tweets that were extracted are shown in Fig. 6. Tweets containing well-known place names such as Kyoto and Kawaramachi could be acquired by either of the two methods. On the other hand, in the optimistic approach, we found that we could also accurately collect objects with relatively small numbers, such as more obscure place names. From these examples, we also confirmed that a keyword-based approach is not always suitable for the purpose. For example, if we set a keyword to "Kyoto," only three tweets were selected, but the other six tweets were not selected and processed.

5. Conclusion

In this paper, we proposed a method for filtering twitter streams using both crowdsourcing and machine learning. In this research, we investigated a problem that occurs when active learning is per-

Output by both optimistic and pessimistic methods.

- It is difficult to understand how to transfer from Kyoto to Hankyu Kawaramachi because you have to walk in from another station!
- Sanjo Station on the Keihan line is a relatively big station. There are various bookstores and bakeries inside, which is convenient.
- Welcome to Kyoto! I will give you Kyoto sightseeing guidance.

Output by only optimistic method.

- The autumn leaves at Imamiya Shrine are beautiful.
- Higashiyama line is far from here.
- My tea candy is so tasty, I should have bought another bag.

Output by only pessimistic method.

- Although it didn't take long to enter the Kyoto Museum, the exhibition room was very crowded.
- I ate sparrow and quail near Fushimi Inari Shrine.
- Now is the best season to see Osmanthus (beautiful tree).

Fig. 6 Results of tweets by optimistic and pessimistic approaches. These tweets are translated from Japanese.

formed on information filtering with regard to 1) the crowdsourcing workers' ability and 2) the tweets presented to the crowdsourcing workers. Information filtering was performed using machine learning and crowdsourcing so as to determine the accuracy and the cost involved in obtaining relevant tweets.

In the evaluation experiment, we evaluated both the workers on crowdsourcing and the machine-learning based classifier. First, in evaluation experiment 1, performance evaluation of the workers was conducted. In information filtering and information retrieval, the proportion of the necessary information in the input texts is extremely small. Therefore, humans thought that there might be a bias to judge the unnecessary text as necessary. Based on the results of the evaluation experiment, it is impossible to say if there is a relationship between the correct answer rate and the correct answer rate. We found that sufficient performance can be obtained even when the correct answer rate is extremely small.

Next, in experiment 2, we confirmed the number of relevant tweets the system can collect when combining crowdsourcing and fastText, a machine-learning based classifier. We compared two strategies, optimistic and pessimistic, for selecting tweets [which were previously] presented to the crowdsourcing workers and which are useful for improving the accuracy of the classifier. We were able to obtain unfavorable tweets with keywords so that we could show the usefulness of the proposed method.

Future work includes the following. We plan to combine the existing keyword-based approach with our proposed crowdsourcing and machine-learning based approach for constructing a more accurate information filtering syste. The information filtering method based on keywords has been proposed and implemented for practical use. The advantage of this approach is that by setting the appropriate keywords, the system can extract relevant tweets at high speed with low cost. The disadvantage is that it is difficult to set the appropriate keywords, and the users cannot input the surrounding keywords related to the appropriate keywords, such as abbreviated keywords and synonyms. On the other hand, although we do not need to set appropriate keywords in this research. Another disadvantage is that a large amount of manual annotation is necessary. Therefore, we should develop an information filtering method by integrating these two methods, in order to collect relevant tweets at low cost.

Coverage is another problem of our proposed method. In our method, the filter may not collect certain tweets which are relevant to a specific topic because there are no similar tweets. To solve this problem, we should collect random tweets from the tweets that were marked as irrelevant by the classifier.

Personalization is also one of our future tasks. When using the proposed method in personalized information filtering, the users need to annotate many tweets. However, it is difficult to obtain a sufficient amount of annotations by one user. Therefore, it is necessary to develop a method to apply efficient collaboration filtering techniques to the proposed method.

Acknowledgments I would like to thank Prof. Satoshi Nakamura for supporting this research project. This work was partly supported by JSPS KAKENHI Grant Number 18H03342, "Research and Development on Fundamental and Utilization Technologies for Social Big Data," the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN, and NAIST Bigdata Project.

References

[1] Suzuki, Y. and Nakamura, S.: Information Filtering Method for Twitter Streaming Data Using Human-in-the-Loop Machine Learning, *Database and Expert Systems Applications* (Hartmann, S., Ma, H., Hameurlain, A., Pernul, G. and Wagner, R. R., eds.), Cham, Springer International Publishing, pp. 167–175 (2018).

[2] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M. and Etzioni, O.: Open Information Extraction from the Web, *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 2670–2676 (online), available from ⟨http://dl.acm.org/citation.cfm?id=1625275.1625705⟩ (2007).

[3] Belkin, N. J. and Croft, W. B.: Information Filtering and Information Retrieval: Two Sides of the Same Coin?, *Commun. ACM*, Vol. 35, No. 12, pp. 29–38 (online), DOI: 10.1145/138859.138861 (1992).

[4] Witten, I. H., Frank, E. and Hall, M. A.: *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition (2011).

[5] Sheng, V. S., Provost, F. and Ipeirotis, P. G.: Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA, ACM, pp. 614–622 (online), DOI: 10.1145/1401890.1401965 (2008).

[6] Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V. and Holzinger, A.: Computational Approaches for Mining User's Opinions on the Web 2.0, *Inf. Process. Manage.*, Vol. 50, No. 6, pp. 899–908 (online), DOI: 10.1016/j.ipm.2014.07.005 (2014).

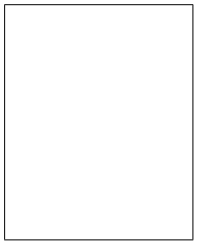
[7] Zaidan, O. F. and Callison-Burch, C.: Crowdsourcing Translation: Professional Quality from Non-professionals, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 1220–1229 (online), available from ⟨http://dl.acm.org/citation.cfm?id=2002472.2002626⟩ (2011).

[8] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R. and Shahabi, C.: Big Data and Its Technical Challenges, *Commun. ACM*, Vol. 57, No. 7, pp. 86–94 (online), DOI: 10.1145/2611567 (2014).

[9] Abel, F., Hauff, C., Houben, G.-J., Stronkman, R. and Tao, K.: Twitcident: Fighting Fire with Information from Social Web Streams, *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, New York, NY, USA, ACM, pp. 305–308 (online), DOI: 10.1145/2187980.2188035 (2012).

[10] Shardanand, U. and Maes, P.: Social Information Filtering: Algorithms for Automating &Ldquo;Word of Mouth&Rdquo;, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co., pp. 210–217 (online), DOI: 10.1145/223904.223931

(1995).
[11] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M.: Short Text Classification in Twitter to Improve Information Filtering, *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, New York, NY, USA, ACM, pp. 841–842 (online), DOI: 10.1145/1835449.1835643 (2010).
[12] Hannon, J., Bennett, M. and Smyth, B.: Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches, *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, New York, NY, USA, ACM, pp. 199–206 (online), DOI: 10.1145/1864708.1864746 (2010).
[13] Uysal, I. and Croft, W. B.: User Oriented Tweet Ranking: A Filtering Approach to Microblogs, *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, New York, NY, USA, ACM, pp. 2261–2264 (online), DOI: 10.1145/2063576.2063941 (2011).
[14] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781 (online), available from ⟨http://dblp.uni-trier.de/db/journals/corr/corr1301.htmlabs-1301-3781⟩ (2013).
[15] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, USA, Curran Associates Inc., pp. 3111–3119 (online), available from ⟨http://dl.acm.org/citation.cfm?id=2999792.2999959⟩ (2013).
[16] Mikolov, T., Yih, W.-t. and Zweig, G.: Linguistic Regularities in Continuous Space Word Representations., *HLT-NAACL*, pp. 746–751 (2013).
[17] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association of Computational Linguistics*, Vol. 5, pp. 135–146 (online), available from ⟨http://aclweb.org/anthology/Q17-1010⟩ (2017).
[18] Xue, B., Fu, C. and Shaobin, Z.: A Study on Sentiment Computing and Classification of Sina Weibo with Word2Vec, *Proceedings of the 2014 IEEE International Congress on Big Data*, BIGDATACONGRESS '14, Washington, DC, USA, IEEE Computer Society, pp. 358–363 (online), DOI: 10.1109/BigData.Congress.2014.59 (2014).
[19] Rocchio, J. J.: Relevance feedback in information retrieval, *The Smart retrieval system: experiments in automatic document processing* (Salton, G., ed.), Prentice Hall, pp. pp. 313–323 (1971).
[20] Grady, C. and Lease, M.: Crowdsourcing Document Relevance Assessment with Mechanical Turk, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 172–179 (online), available from ⟨http://dl.acm.org/citation.cfm?id=1866696.1866723⟩ (2010).
[21] Alonso, O. and Baeza-Yates, R.: Design and Implementation of Relevance Assessments Using Crowdsourcing, *Advances in Information Retrieval* (Clough, P., Foley, C., Gurrin, C., Jones, G. J. F., Kraaij, W., Lee, H. and Mudoch, V., eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 153–164 (2011).
[22] Jörger, P., Baba, Y. and Kashima, H.: Learning to Enumerate, *Artificial Neural Networks and Machine Learning – ICANN 2016* (Villa, A. E., Masulli, P. and Pons Rivero, A. J., eds.), Cham, Springer International Publishing, pp. 453–460 (2016).
[23] Ertekin, S., Huang, J., Bottou, L. and Giles, L.: Learning on the Border: Active Learning in Imbalanced Data Classification, *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, ACM, pp. 127–136 (online), DOI: 10.1145/1321440.1321461 (2007).
[24] Zhang, X., Yang, T. and Srinivasan, P.: Online Asymmetric Active Learning with Imbalanced Data, *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, ACM, pp. 2055–2064 (online), DOI: 10.1145/2939672.2939854 (2016).
[25] Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T.: Bag of Tricks for Efficient Text Classification, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, pp. 427–431 (online), available from ⟨http://aclweb.org/anthology/E17-2068⟩ (2017).
[26] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. and Mikolov, T.: FastText.zip: Compressing text classification models, (online), available from ⟨https://arxiv.org/abs/1612.03651⟩ (2016).
[27] Campbell, C., Cristianini, N. and Smola, A. J.: Query Learning with Large Margin Classifiers, *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 111–118 (online), available from ⟨http://dl.acm.org/citation.cfm?id=645529.657959⟩ (2000).
[28] Lewis, D. D. and Gale, W. A.: A Sequential Algorithm for Training Text Classifiers, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, New York, NY, USA, Springer-Verlag New York, Inc., pp. 3–12 (online), available from ⟨http://dl.acm.org/citation.cfm?id=188490.188495⟩ (1994).
[29] Tong, S. and Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification, *J. Mach. Learn. Res.*, Vol. 2, pp. 45–66 (online), DOI: 10.1162/153244302760185243 (2002).



Yu Suzuki received his M.E. and Ph.D. degree from Nara Institute of Science and Technology in 2001 and 2004, respectively. He became an assistant professor at Ritsumeikan University in 2004, a researcher at Kyoto University in 2009, and an assistant professor at Nagoya University in 2010. He is currently an associate professor at Nara Institute of Science and Technology. His current research interests include Social Web analysis and data mining. He is a member of IPSJ, IEICE, DBSJ, IEEE-CS, and ACM.