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

Yu Suzuki(✉) and Satoshi Nakamura

Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 6300192, Japan
ysuzuki@is.naist.jp

Abstract. There are a massive amount of texts on social media. However, only a small portion of these texts is informative for a specific purpose. If we accurately filter the texts in the streams, we can obtain useful information in real time. In a keyword-based approach, filters are constructed using keywords, but selecting the appropriate keywords to include is often difficult. In this work, we propose a method for filtering texts that are related to specific topics using both crowdsourcing and machine learning based text classification method. In our approach, we construct a text classifier using FastText and then annotate whether the tweets are related to the topics using crowdsourcing. In this step, we consider two strategies, optimistic and pessimistic approach, for selecting tweets which should be assessed. Then, we reconstruct the text classifier using the annotated texts and classify them again. We assume that if we continue instigating this loop, the accuracy of the classifier will improve, and we will obtain useful information without having to specify keywords. Experimental results demonstrated that our proposed system is effective for filtering social media streams. Moreover, we confirmed that the pessimistic approach is better than the optimistic approach.

1 Introduction

A massive amount of texts are on social network services, and much 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 [1]. Information filtering [2] is a process for retrieving texts from text streaming, and the keyword-based method is often used for this. However, it is difficult to set appropriate keywords that correspond to the information needed, because of two reasons; information need are generally vague and sentences on social networks are always broken. Crowdsourcing is one possible solution to solve these issues. However, assessing all tweet streaming data by crowdsourcing is unrealistic, because there are a large amount of tweets, then we should pay many wages to the workers for assessing all tweets. To reduce the amount of crowdsourcing tasks, we used machine learning.

S. Hartmann et al. (Eds.): DEXA 2018, LNCS 11030, pp. 167–175, 2018.
https://doi.org/10.1007/978-3-319-98812-2_13

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

To achieve our objective, we combine two techniques; crowdsourcing and machine learning, which is called Human-in-the-Loop. In this research, we investigated a problem that occurs when active learning is performed on information filtering with regard to the tweets presented to the workers. Information filtering was performed using machine learning and crowdsourcing so as to determine the accuracy and the cost involved in obtaining relevant tweets.

2 Related Work

In social network services, numerous useful texts are posted, such as those related to the spread of influenza or the occurrence of an accident. Several systems have been proposed to capture these incidents quickly [3] and thereby to utilize social media services as a kind of social sensor. Many information filtering methods to glean useful information from streaming data have been proposed, and many of them are used in systems generated for personalization [4] 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 [5], the relationship between users [6], and user behaviors such as re-tweeting [7] are used as features to filter tweets.

Relevance feedback is important for improving the accuracy of information filtering. Rocchio [8] proposed a relevance feedback mechanism on the vector space model. In this mechanism, since more accuracy is improved as more feedback is given, a method using crowdsourcing for feedback has been proposed. [9,10]. However, in these methods, it is assumed that the set of documents that are compatible is sufficiently large as compared with the whole set, so it was not clear whether it can be used for information filtering. In this research, we clarify whether relevance feedback by crowdsourcing is effective when the relevant document is extremely few.

3 Information Filtering Method

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

3.1 Building Classifier

We use FastText [11,12], an application for word embedding and classification, for classifying tweets. With FastText, we input a training set, a pre-trained embedding model, and parameters.

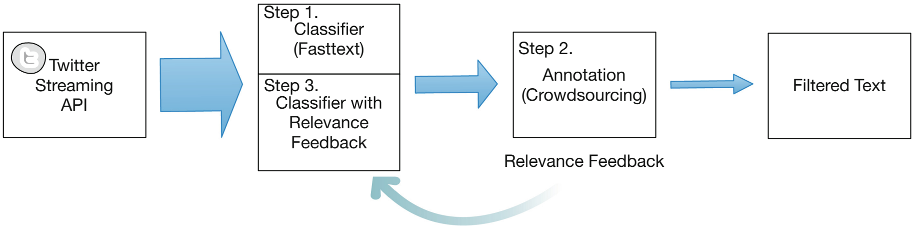
Twitter Streaming API
Step 1. Classifier (Fasttext)
Step 3. Classifier with Relevance Feedback
Step 2. Annotation (Crowdsourcing)
Filtered Text
Relevance Feedback

Fig. 1. Overview of the proposed system.

First, we prepare a dataset for constructing an initial classifier. We remove tweets from the dataset if they include URLs or mention other users. Then, we prepared a set of tweets that are related to specific topics using crowdsourcing. Each tweet is given the label *positive* or *negative*. The label *positive* means that the tweets are relevant to the topics, and the label *negative* means that the tweets are irrelevant to the topics. At this time, the classifier calculates the certainty.

Next, we extract bag-of-word features from the texts. MeCab[1] with the IPADIC-Neologd dictionary[2] was used as a morphological analyzer. There are typically many newly coined words in the tweets, so we use a dictionary that includes new words. Then, we extract nouns, verbs, adjectives, and adverbs. We use a pretrained embedding model[3] constructed using the Japanese Wikipedia copus[4]. This enables synonyms to be handled in the same way.

Finally, using the classifier included in FastText implementation [13], we construct a classifier of the tweets. We assigned a *positive* label or a *negative* label to the unlabeled tweets.

3.2 Classify Tweets

Next, we classify the tweets obtained from text streaming by using the classifier described in Sect. 3.1. We classify the new tweets and continue classification until we can look for tweets labeled as *positive*. The tweets judged as positive by the classifier perform annotation, which is the next step. Tweets that are judged as negative are discarded after this step.

3.3 Annotation

As shown in Fig. 2, we manually judged whether or not the tweets classified as positive by the classifier were positive. 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

[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/.

the accuracy is low. Each judgment made was considered final (i.e., there was no redundancy by multiple workers or averaging that took place).

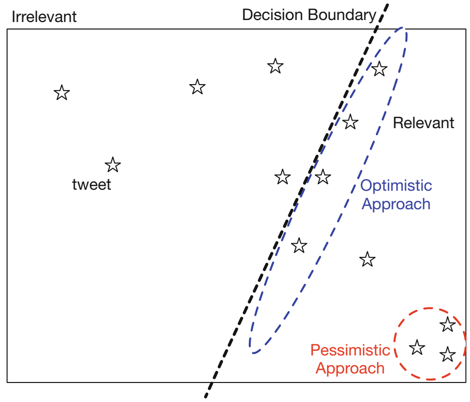

Fig. 2. Intuitive example of our proposed method. (Color figure online)

Interestingly, the performance of the classifier changed depending on which text was judged by a worker. Campbell et al. [14] pointed out that there are differences in the performance of classifiers depending on the sample selection method. The following two ideas were considered as an optimal approach and a pessimistic approach.

Strategy 1: Optimistic Approach. When a classifier classifies a tweet, it is a way to prioritize tweets that are difficult to judge as positive or negative. Lewis et al. [15] proposed a technique called uncertainty sampling, which Simon et al. [16] later used in an application using support vector machines. In this method, tweets near a decision boundary are annotated. The blue part in Fig. 2 shows the tweets which should be annotated using this approach.

Strategy 2: Pessimistic Approach. In this strategy, a presentation is made to the worker in descending order of probability of what the classifier judges as positive. In the information filtering handled in this research, many tweets are considered to be unrelated, which means that even if the positive probability of the classification result is high, it is unlikely to be positive. Therefore, this method was devised to judge, as accurately as possible, the appropriate tweet. The red portion in Fig. 2 shows 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 tweet.

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

Table 2. Experimental setting.

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

4 Evaluation

We used the optimistic and pessimistic approaches (Sect. 3.3) to determine the effectiveness of these strategies using the number of adequate tweets to be obtained (Table 2).

Data. We prepared two groups of tweet data: labeled and unlabeled. As labeled data, we used all 3,580 manually collected tweets. As unlabeled data, the Twitter Streaming API was used to collect more than 1 million tweets in advance. To ensure the same conditions when making a comparison of the two strategies, the tweets to be categorized were made identical for both strategies. If the performance was the same, the same tweet was extracted.

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 positive tweets.
- Classify tweets against positive tweets by crowdsourcing.
- Add judgment result of crowdsourcing to labeled data and return to 1.

First, we gathered tweets by using the optimistic strategy and then collected tweets again by the pessimistic strategy. The number of workers was the same in both strategies, but the workers themselves were different.

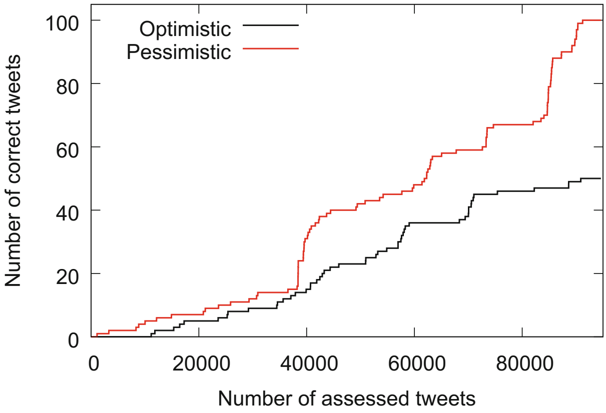

Fig. 3. Ratio of positive tweets and correct answer rate.

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 positive and negative with high accuracy were obtained by grid search.

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 assessments by the pessimistic approach. The results are shown in Fig. 3. We discovered that the correct tweet could be collected twice as fast by the pessimistic approach than 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 4 shows the number of model reconstruction (steps) vs. the ratio of positive tweets. In this figure, a point shows how much percentage of new tweets classifier judged as positive at a step. For example, if the value at step 3 is 0.1, the classifier judges the tweets as positive 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 positive in the first three steps, but after step 4 the ratio is lower than the optimistic approach and the ratio is converged to about 0.02. On the other hand, in the optimistic approach, in the first step, the classifier judge a small number of the positive tweet than that by the pessimistic approach. However, the ratio of positive tweets does not decrease in subsequent steps. The same number of positive tweets are selected at each level. Therefore, decreasing the ratio of positive 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 tweets as compared with the optimistic approach.

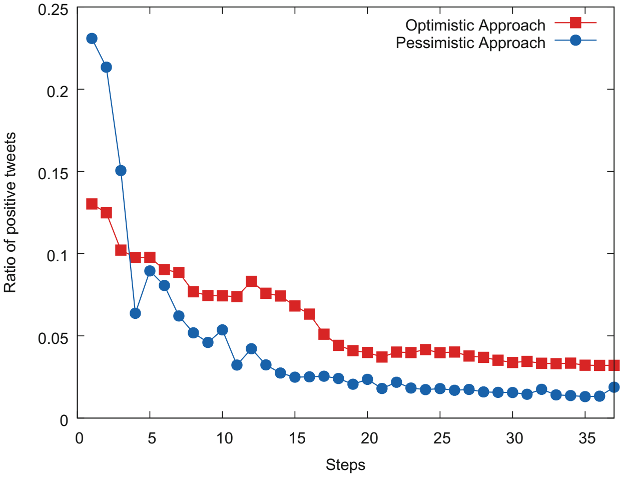

Fig. 4. The number of steps vs. ratio of positive tweets

5 Conclusion

In this paper, we proposed a method for filtering tweet streams using both crowdsourcing and machine learning. In this research, we investigated a problem that occurs when active learning is performed on information filtering with regard to the tweets presented to the 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 confirmed how many relevant tweets the system can collect when combining crowdsourcing and FastText, a machine learning based classifier. At this time, we compared two strategies, optimistic and pessimistic approach, to select tweets presented to workers which are useful for improving the accuracy of the classifier. As a result, we were able to collect relevant tweets as much as 50 tweets and 100 tweets for optimistic approach, and pessimistic approach for 4,500 JPY (45 USD), respectively. At this time, we got an assessment result of around 95,000. We were able to obtain unfavorable tweets with keywords so that we could show the usefulness of the proposed method.

In future work, we should combine the existing keyword-based approach with our proposed crowdsourcing and machine learning based approach for constructing more accurate information filtering.

Acknowledgments. The research results have been achieved by “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.

References

1. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI 2007, pp. 2670–2676, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2007)
2. Belkin, N.J., Croft, W.B.: Information filtering and information retrieval: two sides of the same coin? Commun. ACM **35**(12), 29–38 (1992)
3. Abel, F., Hauff, C., Houben, G.J., Stronkman, R., Tao, K.: Twitcident: fighting fire with information from social web streams. In: Proceedings of the 21st International Conference on World Wide Web, WWW 2012 Companion, pp. 305–308, New York. ACM (2012)
4. Shardanand, U., Maes, P.: Social information filtering: algorithms for automating &ldquo;word of mouth&rdquo;. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1995, pp. 210–217, New York. ACM Press/Addison-Wesley Publishing Co. (1995)
5. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2010, pp. 841–842, New York. ACM (2010)
6. Hannon, J., Bennett, M., Smyth, B.: Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 199–206, New York. ACM (2010)
7. Uysal, I., Croft, W.B.: User oriented tweet ranking: a filtering approach to microblogs. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 2261–2264, New York. ACM (2011)
8. Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) The Smart Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall (1971)
9. Grady, C., Lease, M.: Crowdsourcing document relevance assessment with mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT 2010, pp. 172–179, Stroudsburg, PA, USA. Association for Computational Linguistics (2010)
10. Alonso, O., Baeza-Yates, R.: Design and implementation of relevance assessments using crowdsourcing. In: Clough, P., et al. (eds.) Advances in Information Retrieval, pp. 153–164. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_16
11. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
12. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427–431. Association for Computational Linguistics (2017)
13. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: compressing text classification models, December 2016
14. Campbell, C., Cristianini, N., Smola, A.J.: Query learning with large margin classifiers. In: Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000, pp. 111–118, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2000)

15. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR 1994, pp. 3–12. Springer, New York (1994). https://doi.org/10.1007/978-1-4471-2099-5_1
16. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. **2**, 45–66 (2002)