2F4-OS-5b-04

The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

# Filtering of Impertinent Remarks using distributed expression

Kentaro Hiraishi Daichi Shibata Tomohiro Nishida Naoko Yamaguchi

Shota Suzuki Kai Yoshino Ahmed Moustafa Takayuki Ito

Nagoya Institute of Technology

This paper proposes a filtering method, using supervised learning with distributed representation of posted documents and vectors as features. In recent years, online discussion platforms have witnessed a great popularity. However, there here are many harmful contents such as unrelated spam in these discussion platforms, and violent remarks that insult and discriminate against opponents. As result, it becomes necessary to build a discussion platform that allows online users to participate safely by removing inappropriate remarks. To remove inappropriate remarks, understanding and classifying the meanings of documents is needed. Toward this end, we adopt doc2vec and ELMo to word embedding documents. In addition, we constructed a vectorized document by using document similarity calculation and deep neural networks (DNN). The experimental results show that the proposed method is able to classify with higher accuracy.

## 1. Introduction

Recently, a communication tool allowing users to freely read and write such as bulletin board system and social networking services (SNS) has been developed very much on the web, and everyone can transmit every kind of information on the web. It can be obtained. As a result of on these sites, remarks that disapprove users browsing users such as dating system slandering contents, sexual contents, or speech that encourages crime are scattered, and these contents it is that negative effects are given to minors. And, it is a problem that users who deliver adverse affected contents and users of minors contact each other over the Web, and as a result of recent SNS growth, it is steadily growing. Therefore, it is important to discriminate and remove harmful contents on SNS and Web so as not to adversely affect users of minors.

In recent years, researches on learning meaning vectors of words from large-scale corpus have been actively conducted. In this paper as well, a document is represented by a semantic vector of words contained in the document, and a vector is taken as the feature of the document. We construct information filter using supervised learning, perform evaluation experiment and confirm the effectiveness of this method by showing high F-measure.

## 2. Related Research

Bayesian filtering is a technique used for spam filtering. We learn how a word characteristically appears in spam or non-spam, and perform filtering by calculating the proportion of characteristic words contained in the mail.

In Japanese, Ando et al. [Ando 2010] propose harmful document determination using word collocation information. Prepare a cooccurrence dictionary and calculate harmfulness based on the number of co-occurrence in the positive example and the number of co-occurrence in the negative example.

Otsuka et al. [Otsuka 2014] propose filtering by making the data of the cooccurrence dictionary very large. But as a problem, the execution time takes quite a lot of time.

Sato et al. [Sato 2014] propose filtering using Paragraph Vector and multilayer perceptron. Sato et al. extend Paragraph Vector and predict surrounding words from document vectors and words as input.

## 3. Building Filters

### 3.1 Existing Research

**・Long Short-Term Memory (LSTM)**

Long Short-Term Memory (LSTM) network is one kind of Recurrent Neural Network (RNN) that can learn long-term dependency relationship. RNN has a loop in the intermediate layer in order to retain the storage of input data. Therefore, the neural net can use the previous data as a judgment material. Since the middle layer can deal with the dependence of the data one level before, naturally, when dealing with the previous data, it should be able to deal with the dependency with the first data, but actually It is a problem such as learning can't be done well, there have never been used so much.

LSTM specializes in extending RNN and dealing with previous information.

The choice of information at each gate is made by sigmoid function. LSTM usually learns dependence relationship in forward or forward order, but in this research Bi-LSTM which can learn bi-directionally is used.

**・Gated Recurrent Unit (GRU)**

Gated Recurrent Unit (GRU) is a model that makes LSTM a bit simpler. By introducing forgetting/updating gates like LSTM, it is possible to generate a shortcut path that bypasses the time steps to write, making it easier to maintain the memory of features of events before a long step.

**・doc2vec**

Paragraph Vector [Le 2014], also known as doc2vec, is a vector expression method of documents proposed by Le et al. There are two models, Distributed Memory Model of Paragraph Vectors (PV-DM) and Distributed Bag of Words version of Paragraph Vector (PV-DBOW). In PV-DM, words are predicted from words and document vectors in documents. Fig1 is a schematic diagram of PV-DM

Doctor of Information Science, Nagoya Institute of Technology
Gokiso, Showa, Nagoya, Aichi 466-8555, Japan
E-mail: hiraishi.kentaro@itolab.nitech.ac.jp

In PV-DBOW, predicting words from document vectors. Fig 2 is a schematic diagram of PV-DBOW

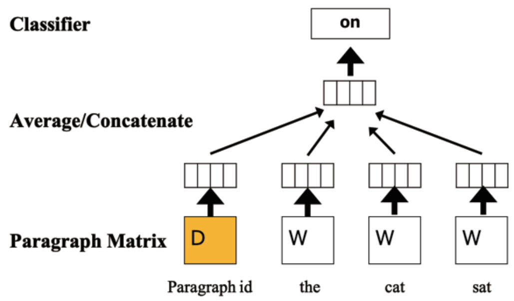


Fig 1: PV-DM

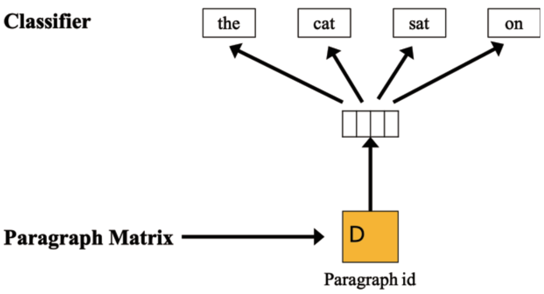


Fig 2:PV-DBOW

## 3.2 Morphological analysis

In this research, MeCab [Kudo 2004] is used as a morphological analysis engine. In order to vectorize documents in Japanese it is necessary to input to dov2vec etc. for each morpheme.

## 3.3 Filtering with ELMo and Bi-GRU and Bi-LSTM

In this method, ELMo is used to vectorize each morpheme. ELMo [Matthew 2018] developed by Matthew et al. is a method to acquire a word expression method considering the context by preliminary learning. This method is realized by acquiring the bidirectional language model (biLM) using a large corpus.

Then build and classify a neural network from the vectorized document. We used Bi-LSTM and Bi-GRU

As indicated in Fig3, Morphologically analyze sentences using MeCab and convert each morpheme into distributed representation with ELMo. Pass each output of the output layer as an input to the layer of LSTM or GRU in order after conversion. Softmax is used as the activation function of the output layer, and the probability of each class can be obtained by this.

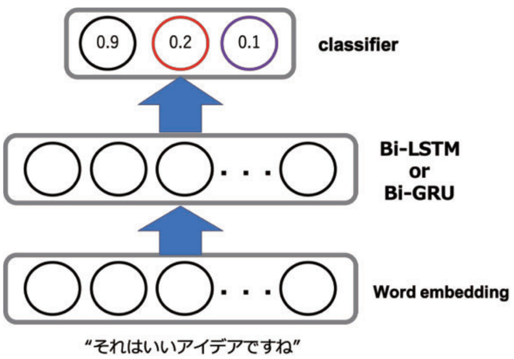


Fig 3: Bi-LSTM and Bi-GRU

## 3.4 Filtering based on document similarity

We propose classification method using cosine similarity. It is necessary to vectorize documents to compute sentence similarity. We use doc2vec to vectorize documents.

The values of each dimension vectorized by doc2vec have an abstract meaning. Therefore, similar sentences can be thought of as having vectors of close angles. We learned sentences of each class and generated feature vectors.

In this method, filtering is performed using doc2vec and cosine similarity. A labeled and vectorized document is used as input to generate a representative vector. We classify newly posted sentences with feature vectors of each class as follows.

$$\arg\max_{c\in C} cos(\vec{q}, \vec{d_c}) = \frac{\sum_{i=1}^{|V|} q_i^2 d_{c_i}^2}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \cdot \sqrt{\sum_{i=1}^{|V|} d_{c_i}^2}}$$

Where, C is the whole class, c is the classification class, $d_c$ is the representative vector of each class, and q is the vector representation of the newly posted document. The similarities calculated for each class are compared and classified into the class with the highest value.

In addition, we propose classification by ensemble in order to increase accuracy. The ensemble method is a method of improving prediction accuracy by fusing classifiers that have separately learned separately.

$$\arg\max_{c\in C} \frac{1}{N}\sum_{k=1}^{N} \cos(\vec{q}, \vec{d}_{c_k})$$

As shown in equation, in this method, multiple models of doc2vec are created, In the model of cosine similarity with new posting data was calculated and averaged to classify it to the highest one. As shown in equation, in this method, multiple models of doc2vec are created, In the model of cosine similarity with new posting data was calculated and averaged to classify it to the highest one.

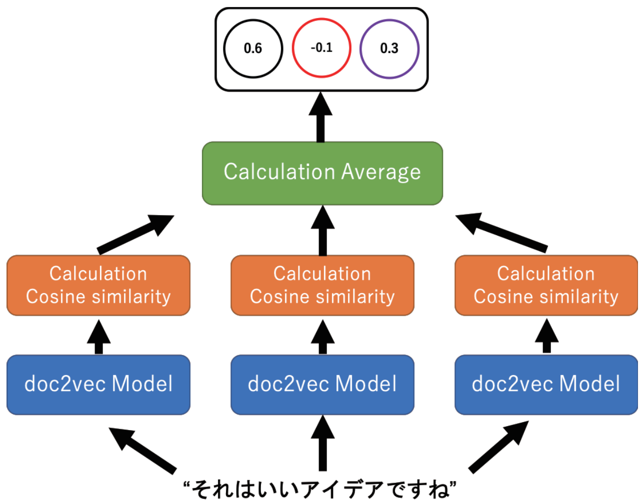


Fig 4: doc2vec Architecture

## 4. Evaluation Experiment

### 4.1 Experimental Design

In this experiment, we gathered about 10,000 data from a Japanese textboard, COLLAGREE [Ito 2015] and D-Agree operating in this laboratory. Evaluate with 7,520 training data and 1,680 test data.

As a proposed method, experiments are performed using doc2vec as an ensemble model, Bi-LSTM model, and Bi-GRU model, and the usefulness of the proposed method is confirmed by obtaining high evaluation index. In this research, as the evaluation index of the classification method, the average value of F-measure when classifying from test data was used. F-measure is the harmonic mean of the relevance rate and recall rate.

### 4.2 Results and Discussion

The average value of F-measure for classification using document similarity and classification using Bi-RNN is shown in the following Table 1.

Tab 1: Experiment Results

| Method | F-measure |
|---|---|
| doc2vec Ensemble Classifier | 0.9360 |
| Bi-LSTM | 0.9190 |
| Bi-GRU | 0.9164 |

Results on Precision, Recall and F-measure in each method are shown. Table 2 shows results in none-toxic document. Table 3 shows results in obscene document. Table 4 shows results in violent document.

Tab 2: Results in None-toxic Document

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| doc2vec Ensemble Classifier | 0.9306 | 0.9882 | 0.9585 |
| Bi-LSTM | 0.9284 | 0.9450 | 0.9366 |
| Bi-GRU | 0.9022 | 0.9701 | 0.9349 |

Tab 3: Results in Obscene Document

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| doc2vec Ensemble Classifier | 0.9600 | 0.9041 | 0.9312 |
| Bi-LSTM | 0.9340 | 0.9386 | 0.9363 |
| Bi-GRU | 0.9545 | 0.9144 | 0.9340 |

Tab 4: Results in Violent Document

| Method | Precision | Recall | F-measure |
|---|---|---|---|
| doc2vec Ensemble Classifier | 0.9310 | 0.9060 | 0.9183 |
| Bi-LSTM | 0.9013 | 0.8670 | 0.8839 |
| Bi-GRU | 0.9176 | 0.8460 | 0.8803 |

In the approximate class, ensemble classification by doc2vec showed the highest F-measure.

As shown in the Table1, the ensemble model using doc2vec shows the highest F-measure. Because, there was a difference in recall and precision between models used for the ensemble. By ensemble, the balance between precision and recall got better, leading to higher F-measure than other methods.

## 5. Conclusions and Future work

This paper proposes a method to classify whether posting is inappropriate by using distributed representation. In the proposed method, documents are vectorized using doc2vec, representative vectors are generated for each class, and classified by cosine similarity. Also, we vectorize each morpheme using ELMo and classify it using Bi-RNN. In order to evaluate the proposed method, we conduct experiments using post data on online textboard. The experimental results demonstrate that the proposed method classifies inappropriate posts with high accuracy. Future work is planned to subdivide classes of inappropriate submissions.

### Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR15E1, Japan.

### References

[Ando 2010] Satoshi Ando, Yutaro Fujii, Takayuki Ito. Filtering Harmful Sentences based on Multiple Word Co-occurrence. Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on. IEEE, 2010.

[Otsuka 2014] Takanobu Otsuka, Deyue Deng, Takayuki Ito. Text Filtering for Harmful Document Classification Method Using Three words Co-occurrence and Large-scale Data Processing. IEEJ Transactions on Electronics, Information and Systems 134.1: pp168-175, 2014

[Sato 2014] Genki Sato, Takayuki Ito. A Harmful Document Classification Method Using Paragraph Vector. JAWS2014: pp321-324, 2014

[Le 2014] Quoc V. Le, Tomas Mikolov. Distributed Representations of Sentences and Documents. JMLR W&CP 32 (1): pp.1188-1196, 2014.

[Matthew 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. Deep contextualized word representations. NAACL 2018, 2018.

[Kudo 2004] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP- 2004), pp.230-237, 2004.

[Ito 2015] Takayuki Ito, et al. Incentive Mechanism for Managing Large-Scale Internet-Based Discussions on COLLAGREE. Collective Intelligence 2015, 2015.