International Journal of Advanced Intelligence
Volume 10, Number 1, pp.121-133, March, 2018.

AIA International Advanced Information Institute

Emotion Estimation Adapted to Gender of User Based on Deep Neural Networks

Naoya Fujino, Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita
Tokushima University, Minamijosanjima-cho 2-1, Tokushima
Tokushima, 770-8506, Japan
matumoto@is.tokushima-u.ac.jp
mino@is.tokushima-u.ac.jp
kita@is.tokushima-u.ac.jp

Received (15 Sep. 2017)
Revised (15 Dec. 2017)

In this study, we focus on Twitter as a representative SNS and target emotion estimation from tweets posted on Twitter by male and female users. Specifically, we construct gender-based emotion estimation models assuming that there are different word usage tendencies between genders. By analyzing gender-specific differences in the use of emotion-related slang and emoji, we propose a method to improve emotion estimation based on neural networks using a different distributed representation model for each gender. Our evaluation experiments show that training with Deep Convolutional Neural Networks using word's distributed representation as the feature produced higher estimation accuracy than training with Feed Forward Neural Networks.

Keywords: emotion estimation, user's gender, deep neural networks

1. Introduction

With the growth of the Internet, users with various attributes routinely use bulletin boards, blogs, SNS, etc. to communicate information and exchange opinions. Postings by users typically include contents related to user attributes such as gender and job. It is thus possible to identify collective trends from such contents based on a variety of attributes.

Emotion estimation techniques appear to be highly useful in promoting smooth communication between humans and computers. It would seem to be especially useful, then, to have a technique capable of estimating emotion from texts posted on weblogs and SNS that are used by many users and are updated daily.

In this study, we focus on Twitter [1] as a representative SNS and target emotion estimation from tweets posted by male and female users in order to develop gender-specific emotion estimation models. More concretely, we analyze the relationship between apparent emotional tendencies and gender and formulate emotion estimation models based on these relationships. Building on the assumption that there are different word usage tendencies between genders, we analyze gender-related differences in the use of slang and emoji that are thought to be connected to emotion.

In related work, Ott[2] analyzed by gender tweets written in English and identified several features by using different machine learning methods. We construct sentiment analysis models for male and female Twitter users using a machine learning method and propose a method to improve emotion estimation based on neural networks using a distributed representation model for each gender.

2. Related Works

Volkova et al. [3] conducted sentiment analysis on tweets in Spanish and Russian according to gender and achieved improved accuracy by weighting gender-specific vocabulary using expressions such as hashtags and emoticons as features. Iosub et al.[4] analyzed differences in emotional expressions according to gender by targeting the talk page of the English Wikipedia. However, the target language and research purpose of their work are different from ours.

Iwasa and Matsumoto [5],[6] proposed a method for creating an emotion estimation model by various machine learning methods such as the naive Bayes method according to different attributes. Their work was based on an assumption that user attributes would appear as features in sentences and emotional expressions using words. However, they were unable to produce the expected results because of insufficient training data. On the other hand, increasing the amount of training data would likely increase the annotation cost without ensuring superior results commensurate with the cost increase.

In fact, we believe that the quality of annotation could actually decrease with an increase in the amount of data. Recently, crowdsourcing has become very popular [7],[8],[9],[10]. Crowdsourcing is a process for acquiring necessary input for a task or project by enlisting a large number of contributors. In research, this process is often used for sorting tasks requiring human judgment, such as labeling on a corpus. However, in crowdsourcing, the quality of annotation depends on the annotators. Although quality could be improved by increasing the number of annotators, this would have drawbacks such as increased time and cost, depending on the nature of the task involved. Moreover, even if the amount of data were increased, the problem of unknown expressions would likely remain.

In this paper, we use distributed vector representations of sentences in order to solve these problems and apply the sentence vectors to create gender-related emotion estimation models. Specifically, we use sentence2vec [11] to create our distributed vector representations for sentences. Sentence2vec is based on word2vec[12], the learning algorithm for distributed expression vectors for words suggested by Mikolov et al. The word2vec[12] algorithm was developed into paragraph2vec[13] by vectoring the distributed expressions by sentence unit. Sentence2vec is one of the implementations of paragraph2vec. As with word2vec, when a tokenized text corpus is given as training data, we are able to create a model capable of generating distributed vector representations by sentence unit. To solve the problem of bias due to insufficient training data, we automatically annotate emotion labels to tweets

that were collected by using emoji as query, and use these to form a large set of training data. For comparison, we also apply a method to train the emotion estimation model by convolutional neural networks (CNN) using the word distributed representation vector as the feature.

3. Proposed Method

Iwasa et al. collected 300 tweets from 15 male and 15 female Twitter users (9,000 tweets in total) who were well-known public figures and added annotated emotion labels to the tweets using four annotators. One or more labels were annotated to each tweet by multiple annotators. Table1 shows the emotion labels used in the experiment and the number of sentences that were annotated with the labels. Table2 shows concrete examples of sentences with the annotation of emotion labels. In this study, we redefined emotion labels by regarding four emotions as basic emotions following Fisher's emotional system diagram [?,?].

Table 1. Emotion Labels and Number of Sentences

Joy	Surprise	Anger	Sorrow	Total
8,551	432	275	1,450	10,708

Table 2. Example of Annotation

Emotion	Tweet	Gender
Joy	I love curry pan!!	Male
Surprise	I didn't think you can do a back hip circle.	Female
Anger	I was jerked from slumber by mosquito... I feel itchy...	Male
Sorrow	Sorry, I couldn't go...	Female

3.1. *Emotion Estimation Model Based on Sentence Distributed Representation Vector*

In our study, we focused on 142 popular male and 166 popular female accounts. In all, we collected 26,390 male tweets and 32,619 female tweets to train our models to generate a distributed vector representation for sentences. The accounts of well-known figures were used because their gender information was more apparent than would be the case for general users.

We did not annotate labels on these tweets. Instead, we tokenized them by word unit using a morphological analyzer, then created male and female tweet corpora for training. We created models to generate the distributed representation vector of

sentences by learning sentence2vec with a dimension number of 500 and a context window size of 5, and with mean CBOW.

By using a sentence vector generation model, it is possible to convert tweet sentences into fixed dimension and dense real-valued vectors. With the vector as the feature, we trained an emotion estimator using neural networks (NN), a machine learning method. A certain level of estimation accuracy could be expected for tweet sentences, even those that included unknown expressions. In order to compensate for the lack of training data and to adjust to the gender of the evaluation target, we converted tweets to feature vectors by using a sentence vector generation model for each gender.

3.2. *Emotion Estimation Model Based on Word Distributed Representation Vector*

As described in the previous section, the distributed vector representation of a sentence is based on word distributed vector representations. This subsection proposes an emotion estimation model using word embedding, wherein we construct word embedding models for each gender by training word representation vectors with the respective tweet corpus of male and female users. We then convert emotion labeled sentences into lists of word embedding vectors based on a word embedding vector model.

Sentences can include unknown words not present in the word embedding vector model. We treat these words as "invalid words" and the other words as "valid words." We train the emotion estimation model by inputting the lists of word embedding vectors into deep convolutional neural networks (DCNN).

As shown in Fig.1, we input as the training data for DCNN a 500 x 20 matrix consisting of a 500-dimension word embedding vector of valid words for 20 words.

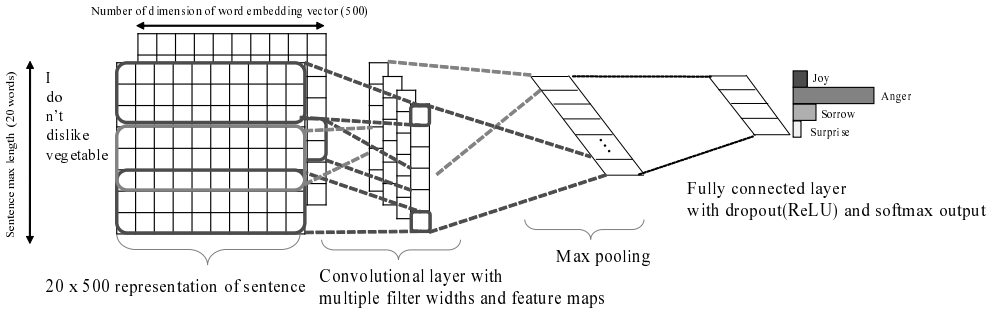

Fig. 1. Training of DCNN

3.3. *Distant Supervision Based on Emoji*

We used a method to automatically annotate an emotion label to unlabeled tweets based on emoji in order to compensate for the lack of extensive training data. Because there are emoji that express emotion, we annotate an emotion label to emoji that are available on Twitter by using these emoji and the emoji data vectorized by Emoji2Vec[15].

Emoji2Vec is the method or tool proposed by Eisner et al. that converts emoji into a distributed representation by training large tweet sentences containing the emoji based on the skip-gram model. We annotated the emotion label to the tweets containing an emoji based on the emoji having emotion labels. We defined this emotion-labeled corpus as the emoji emotion tweet corpus (EETC).

4. Evaluation of Sentence Vector Based Method

4.1. *Experimental Method*

To confirm the effectiveness of the proposed method, we conducted evaluation experiments by cross-validation. In the cross-validation test, we chose the tweets of 30 users (15 males, 15 females) to be used as evaluation data. The tweets of the remaining 29 users were used as training data (30-fold cross validation). The training data were not divided by gender since the number of emotion-labeled tweet sentences was small. A previous study showed that there was no distinct difference in emotion estimation results based on gender but that the influence of the size of the training data set was significant. Therefore, we decided to consider the influence of gender when the distributed representation vector of sentences was generated. We verified estimation accuracy for all users and calculated the average values as the evaluation index for comparison. As shown in Table3, we compared the experimental results combining multiple conditions when the Twitter user of the labeled evaluation tweet was “male” or “female,” when the tweets used to generate the sentence vector were written by either a “male” or “female,” and when both “males” and “females” were users.

Table 3. Experimental conditions

		Gender of Pre-training Tweets		
		Male(M)	Female(F)	Male(M) and Female(F)
Gender of Evaluation Tweets	Male(M)	(a)	(b)	(c)
	Female(F)	(d)	(e)	(f)

We also conducted a comparison experiment to compare normal NN with one hidden layer and a deep neural network consisting of two or more layers (two-layers and three-layers). We used a feedforward neural network as a type of NN and used empirically effective values for the other parameters. Experimental results were

evaluated using accuracy, recall, precision and F1-score. The calculation formula for each index is as Eq.1,2,3,4.

$$Accuracy = \frac{C}{N} \tag{1}$$

$$Recall = \frac{\sum_{(e \in L)} \frac{C_e}{N_e}}{|L|} \tag{2}$$

$$Precision = \frac{\sum_{(e \in L)} \frac{C_e}{O_e}}{|L|} \tag{3}$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

C indicates the number of sentences that the estimated labels matched with the correct labels. N indicates the total number of evaluation sentences. e indicates emotion label, L indicates a set of emotion labels. C_e indicates the number of sentences with correct estimation results when a sentence with correct label as e is input. N_e indicates the number of the sentences with correct label as e. $|L|$ means the varieties of emotion labels (four kinds). O_e indicates the number of sentences whose labels were estimated as e.

4.2. *Experimental Results*

The accuracy of experimental results is shown in Fig.2; recall, precision, and F1-scores are shown in Table4. In Table4, l_n indicates the number of layers.

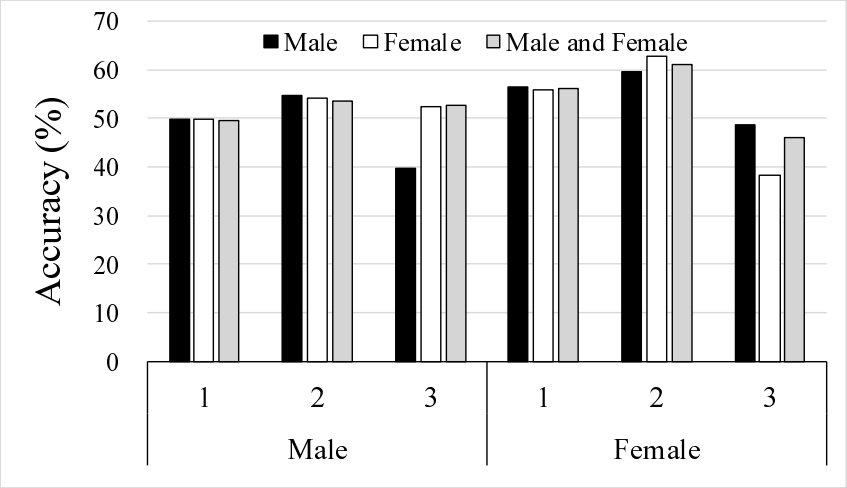

Fig. 2. Accuracy of Experimental Results

Table 4. Experimental Results

	l_n	M				F				M and F			
			R	P	F1		R	P	F1		R	P	F1
M	1	(a)	0.32	0.27	0.24	(b)	0.32	0.27	0.24	(c)	0.31	0.27	0.23
	2		0.32	0.27	0.25		0.31	0.27	0.25		0.32	0.27	0.25
	3		0.25	0.24	0.19		0.26	0.16	0.22		0.25	0.23	0.22
F	1	(d)	0.34	0.28	0.26	(e)	0.34	0.32	0.26	(f)	0.34	0.32	0.26
	2		0.33	0.27	0.27		0.32	0.27	0.27		0.34	0.28	0.27
	3		0.28	0.25	0.24		0.26	0.24	0.18		0.25	0.23	0.20

According to the results, the highest accuracy was obtained under conditions (a) and (e), when the gender of the evaluation tweets is the same as the gender of the pre-trained tweets. In terms of layers, the highest accuracy was obtained when the number of intermediate layers of the NN was set to two.

Furthermore, when the number of intermediate (hidden) layers was set to two, accuracy improved with any combination of the inputted sentences and models. However, this does not mean that increasing the number of hidden layers will necessarily result in greater accuracy. In fact, accuracy drastically decreased with three or more hidden layers. By increasing the number of intermediate layers, more complex learning becomes possible; however, it may be that over-learning occurs due to the complexity of the learning involved.

We found in all models that the estimation accuracy for females was higher, on average, than the estimation accuracy for males. This was possibly because the tweets of female users were richer in emotional expressions than those of male users. In addition, a comparison of the estimation accuracy of each emotion label revealed that the estimation F1-scores of the two labels “surprise” and “anger” were extremely low. The estimation results of the model with the best estimation accuracy are shown for each emotion label in Table5.

Table 5. Estimation F1-score for each emotion label.

Emotion	F1-score
Joy	0.629
Surprise	0.000
Anger	0.000
Sorrow	0.669

As indicated, the F1-score of the emotion labels “surprise” and “anger” was 0%. We believe that estimation accuracy can be improved, at least to a certain extent, by eliminating bias in the number of instances for each emotion label. Furthermore,

it will be necessary to add diversity by not only increasing the size of the corpus but also by expanding the pool of target Twitter users for labelling, from famous celebrities only to general users with gender information in their profiles.

4.3. *Analysis*

Having found gender-related differences in accuracy, we analyzed the expression appearance tendencies in sentences for each gender. Specifically, we performed three types of analysis. In each, the analysis target was the tweets that were used as our evaluation data (written by users of both genders and with the annotation of emotion labels). Table 6,7, and 8 summarize the results of analysis (1), (2), and (3).

(1) Appearance tendency of word's emotion polarity
(2) Appearance tendency of emoji
(3) Appearance tendency of unvectorized expression

Table 6. Appearance tendency of word's emotion polarity

Emotion polarity	Male	Female
Negative	4.44	4.14
Positive	1.89	1.75

Table 7. Appearance Tendency of Emoji

Male	Female
0.031	0.195

Table 8. Appearance tendency of unvectorized expression

Type of pre-training corpus \ Type of evaluation corpus	Male	Female
Male	1.157	1.572
Female	1.197	1.547
Male and Female	1.003	1.347

Significant differences in the tendency of emoji use were found according to gender in all three analyses. Fig.3 shows the emoji usage rate for each user. We found that over 80% of both male and female users used emoji at least once, although there were differences depending on the user. Moreover, we found that four female

users used emoji over 100 times, which meant that they used emoji almost every day. These results suggest that the use of emoji may affect the accuracy of emotion estimation. Clearly, it would be difficult to estimate emotion from tweets that do not contain words expressing emotions. However, we believe that emoji can be an effective substitute for emotion expressions since the emotion estimation accuracy for female users who frequently used emoji was high.

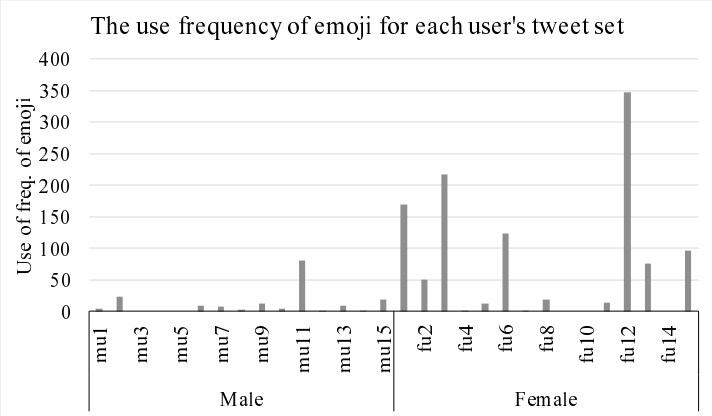

Fig. 3. Frequency of emoji in each user's tweet set

There was little difference in emotion polarity or unvectorized expressions between males and females. However, the number of unvectorized expressions was small in the male corpus and also in the male and female corpus. This tendency was observed in the case of females, as well. It is possible that factors other than gender, such as hobbies or occupation, have an influence.

5. Evaluation of Distant Supervision Based on Emoji

As we found that there were distinct characteristics in the usage tendencies of emoji, we annotated emotion labels based on Distant Supervision on a tweet corpus collected by using emoji as the query. An emotion estimation model was created by training our emotion-labeled corpus with the two hidden layer neural network that obtained the best results in the previous Section. We then conducted emotion estimation experiments on male and female tweets by using the created emotion estimation model. We also conducted another experiment using Deep Convolutional Neural Networks (DCNN), which is a type of deep neural network.

Rather than using distributed representation vectors for sentences, we used word vectors trained with fastText[fastText],[Bojanowski] as the feature. We defined the upper limit of the number of available words as 20. If a sentence contained fewer than 20 words, we applied a padding process to the sentence. The training emotion

labeled corpus based on emoji (EETC) included 113,696 tweets (28,424 tweets for each emotion label).

5.1. *Experimental Results*

The accuracy, recalls, precision and F1-scores produced by each of the emotion estimation models are shown in Fig.4, and Table9, 10.

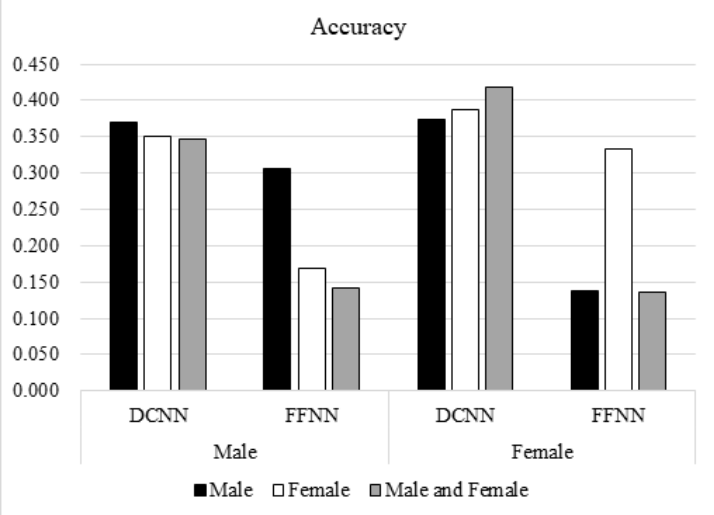

Fig. 4. Accuracy of DCNN and FFNN

Table 9. Result of FFNN with hidden 2-layer (same gender)

	Male			Female		
Emotion	R	P	F1	R	P	F1
Joy	0.163	0.742	0.267	0.16	0.754	0.264
Anger	0.361	0.067	0.113	0.262	0.064	0.102
Sorrow	0.309	0.181	0.229	0.258	0.172	0.206
Surprise	0.263	0.075	0.117	0.393	0.095	0.154
Avg.	0.274	0.266	0.182	0.268	0.271	0.182

Although overall accuracy decreased, balanced estimation was achieved, at least to some extent, because of the higher F1-scores of “anger” and “surprise,” which were greater than 0. One of the main reasons for the failure of F1-scores to improve as a whole may have been that only emoji were used as a clue in annotating labels

Table 10. Result of DCNN using word vector feature (same gender)

	Male			Female		
Emotion	R	P	F1	R	P	F1
Joy	0.778	0.377	0.508	0.762	0.395	0.52
Anger	0.106	0.338	0.161	0.116	0.292	0.166
Sorrow	0.236	0.469	0.314	0.228	0.471	0.307
Surprise	0.064	0.123	0.084	0.07	0.136	0.092
Avg.	0.296	0.327	0.267	0.294	0.324	0.271

to the training data. It would be useful to analyze whether the accuracy level would change if we reduced the influence of emoji on the training data, or we removed emoji from both the test data and the training data.

6. Conclusion

In this paper, we generated sentence/word distributed representations using the tweet corpora of males and females as pre-training corpora to validate the effectiveness of gender-specific emotion estimation models. Although we observed some level of effectiveness, we also found that "surprise" and "anger" were barely estimated due to the bias of the training data. To reduce this bias, we created an emotion estimation model by using other training data that were collected based on emoji. Consequently, the bias in emotion estimation was revised and accuracy was somewhat improved.

However, generally, the reliability of the labels of the collected training data based on emoji was less than that of the manually labeled corpus. Therefore, using labeling rules based on emoji and emotion expressions in sentences would appear to be necessary to increase the reliability of the labels.

Training with DCNN using word distributed representations as the feature produced better estimation accuracy than training with FFNN. We intend to further investigate the reason for this difference in order to determine whether it is due to using sentence distributed representations or to over-fitting by FFNN.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K00425, 15K00309, 15K16077.

References

1. Twitter: https://twitter.com/
2. M. Ott: "Tweet Like a Girl: A Corpus Analysis of Gendered Language in Social Media," bachelor thesis at Yale University, 2016.
3. S. Volkova, T. Wilson, and D. Yarowsky: "Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP2013), 2013.
4. D. Iosub, D. Laniado, C. Castillo, M. F. Morell, and A. Kaltenbrunner: "Emotions under Discussion: Gender, Status, and Communication in Online Collaboration," PLOS ONE, Vol.9, Issue 8, 2014.
5. F., Iwasa, K. Matsumoto, M. Yoshida, and K. Kita: "Emotion Estimation of Twitter for Each Attribute," Annual Meeting of The Association for Natural Language Processing, 2016.
6. K. Matsumoto, M. Yoshida, K. Kita, Y. Wu and F. Iwasa, "Effect of Users Attribute on Emotion Estimation from Twitter," Proceedings of the 2nd IEEE International Conference on Computer and Communications, Vol.3, pp.1186–1190, Oct. 2016
7. Marta Sabou, Kalina Bontcheva, Leon Derczynski, Arno Scharl, "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines," In:Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), (2014).
8. Bontcheva K., Derczynski L., Roberts I., "Crowdsourcing Named Entity Recognition and Entity Linking Corpora," Handbook of Linguistic Annotation, 2017.
9. M Soleymani and M Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010.
10. Jha, Mukund, Andreas, Jacob, Thadani, Kapil, Rosenthal, Sara, McKeown, Kathleen, "Corpus Creation for New Genres: A Crowdsourced Approach to PP Attachment," https://doi.org/10.7916/D8FF41PF, 2010.
11. Sentence2Vec: https://github.com/klb3713/sentence2vec.
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, CoRR, abs/1310.4546 (2013)
13. Mikolov, T.: Distributed Representations of Sentences and Documents, Proceedings of the 31 st International Conference on Machine Learning, JMLR: W&CP Vol.32, (2014)
14. Fischer, K. W., Shaver, P. and Carnchan, P.: A skill approach to emotional development: From basic- to subordinate-category emotions. In W. Damon (Ed.), Child development today and tomorrow, pp.107–136, (1989)
15. Eisner, B., Rocktschel, T., Augenstein, I., Matko Bonjak, M., and Riedel, S.: emoji2vec: Learning Emoji Representations from their Description, In Proceedings of the 4th International Workshop on Natural Language Processing for Social Media at EMNLP (2016)
16. fastText: https://github.com/facebookresearch/fastText.
17. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov : Enriching Word Vectors with Subword Information arXiv preprint arXiv:1607.04606, 2016.

Naoya Fujino

He received bachelor's degree in 2017 at Tokushima University. His research interests include Natural Language Processing, Social Network Analysis, and Deep Learning.

Kazuyuki Matsumoto

He received the PhD degree in 2008 from Tokushima University. He is currently an assistant professor of Tokushima University. His research interests include Affective Computing, Emotion Recognition, Artificial Intelligence and Natural Language Processing. He is a member of IPSJ, ANLP, IEICE and IEEJ.

Minoru Yoshida

He is a lecturer at the Department of Information Science and Intelligent Systems, University of Tokushima. After receiving his BSc, MSc, and PhD degrees from the University of Tokyo in 1998, 2000, and 2003, respectively, he worked as an assistant professor at the Information Technology Center, University of Tokyo. His current research interests include Web Document Analysis and Text Mining for the Documents on the WWW.

Kenji Kita

He received the B.S. degree in mathematics and the PhD degree in electrical engineering, both from Waseda University, Tokyo, Japan, in 1981 and 1992, respectively. From 1983 to 1987, he worked for the Oki Electric Industry Co. Ltd., Tokyo, Japan. From 1987 to 1992, he was a researcher at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. Since 1992, he has been with Tokushima University, Tokushima, Japan, where he is currently a Professor at Faculty of Engineering. His current research interests include multimedia information retrieval, natural language processing, and speech recognition.