J Inf Process Syst, Vol.19, No.5, pp.641~651, October 2023
https://doi.org/10.3745/JIPS.04.0287

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

BERT-Based Logits Ensemble Model for Gender Bias and Hate Speech Detection

Sanggeon Yun, Seungshik Kang, and Hyeokman Kim*

Abstract
Malicious hate speech and gender bias comments are common in online communities, causing social problems in our society. Gender bias and hate speech detection has been investigated. However, it is difficult because there are diverse ways to express them in words. To solve this problem, we attempted to detect malicious comments in a Korean hate speech dataset constructed in 2020. We explored bidirectional encoder representations from transformers (BERT)-based deep learning models utilizing hyperparameter tuning, data sampling, and logits ensembles with a label distribution. We evaluated our model in Kaggle competitions for gender bias, general bias, and hate speech detection. For gender bias detection, an F1-score of 0.7711 was achieved using an ensemble of the Soongsil-BERT and KcELECTRA models. The general bias task included the gender bias task, and the ensemble model achieved the best F1-score of 0.7166.

Keywords
BERT Embedding Model, Gender Bias, Hate Speech Detection, Logistics Ensemble

1. Introduction

Online users express their opinions through short sentences in news comments, online communities, and social network platforms [1-4]. In these environments, opinions are freely expressed because anonymity is guaranteed; however, it has severe side effects: hateful expressions are overused [5-8]. Therefore, we attempted to investigate the severity of cyberbullying and celebrity suicides caused by malicious comments from anonymous users. Major portal companies use a function to remove malicious comments or prevent readers from writing such comments. However, it is difficult to filter malicious comments, as there are diverse ways to express them in words. Recently, a method of classifying malicious comments has been investigated to solve this problem, and hateful expression detection is applied to a news comment filtering system [9-12]. Methods for developing such a system include deep learning and bidirectional encoder representations from transformers (BERT)-based pretraining [13-17]. BERT is a pretrained embedding model that uses a transformer and exhibits good performance in natural language processing. After the advent of BERT, many attempts have been made to develop language models that outperform BERT, for example, transformer-based models such as ALBERT, RoBERTa, and ELECTRA.

Manuscript received September 6, 2022; first revision November 29, 2022; second revision January 4, 2023; accepted January 29, 2023.

* **Corresponding Author:** Hyeokman Kim (hmkim@kookmin.ac.kr)
Dept. of Computer Science, Kookmin University, Seoul, Korea (ggj06281@kookmin.ac.kr, sskang@kookmin.ac.kr, hmkim@kookmin.ac.kr)
Current affiliation for author, Sanggeon Yun, is Dept. of Computer Science, University of California, Irvine, CA, USA.

Regarding the Korean language, BERT-based models have been built on large-scale text corpus and evaluated using various datasets [18,19]. For example, the types of hateful expressions have been analyzed using user comments on news articles. Cho and Moon [20] developed a hate speech and gender bias dataset for machine learning. They made it freely available on GitHub and opened a hate expression detection contest on Kaggle to share model performance and evaluation results. This contest comprises three topics: general bias, gender bias, and hate speech detection. In hate speech detection tasks, attempts have been made to achieve higher performance using the ensemble method. Zimmerman et al. [21] developed a logits ensemble model with equal weights to the result vectors of a single model and achieved an improvement of approximately 2%. In contrast to a logits ensemble, Karim et al. [22] proposed a majority voting-based ensemble and achieved an improvement of approximately 1.15%. In this study, we aim to achieve better performance improvement by proposing a novel logits ensemble model that combines various models using different weighting schemes. We evaluated model performance by participating in the Kaggle competition. The remainder of this paper is organized as follows. Section 2 introduces studies related to pretrained BERT embedding models. The hate speech dataset and the logits ensemble model with a label distribution are described in Section 3. Section 4 presents the experimental results and hyperparameter setting with finetuning. Section 5 describes future work with concluding remarks.

2. Pretrained BERT Embedding Models

To evaluate the performance of our model, four pretrained embedding models are used as base models. Table 1 shows the pretrained models trained independently, that is, they are trained with their datasets.

Table 1. Pretrained embedding models

Model	Pretrained models
BERT-base	$BERT_{KoBERT_{SKT}}$
RoBERTa-base	$RoBERTa_{soongsil-bert}$, $RoBERTa_{KLUE}$
ELECTRA-base	$ELECTRA_{KcELECTRA}$

$\boldsymbol{BERT_{KoBERT_{SKT}}}$

KoBERT (Korean BERT pretrained) is a BERT-based model built by SKT. The SKT NLP group attempted to outperform Google’s multilingual BERT model. Training has been conducted using a dataset consisting of 5 million sentences and 54 million words using Korean wiki text. It has a dictionary of 8,002 tokens trained on Korean wiki text by the SentencePiece tokenizer.

$\boldsymbol{RoBERTa_{soongsil-bert}}$

Soongsil-BERT is a pretraining model based on the RoBERTa model built at Soongsil University. Most Korean BERT models are trained on refined datasets such as the Korean wiki text and news articles. However, this model was built on SNS data of internet communities, a corpus comprising speech sentences and web documents collected by the National Institute of the Korean Language. The dataset consists of more than 100 million sentences, and its dictionary trained by the byte pair encoding (BPE) tokenizer has 16,000 tokens.

$RoBERTa_{KLUE}$

It is a RoBERTa-based pretrained model built for the Korean Language Understanding Evaluation (KLUE) platform. It was developed to provide a baseline model, along with the construction of a dataset for evaluating Korean language models. By combining five corpora, more than 400 million sentences were collected. Its dictionary has 32,000 tokens and is trained by the BPE tokenizer.

$ELECTRA_{KcELECTRA}$

KcELECTRA (Korean comments ELECTRA) is a pretrained model based on the ELECTRA model. Similar to Soongsil-BERT, it was developed to reflect real-world characteristics such as new words and typos. The dataset comprises 100 million sentences with 30,000 tokens collected from the comments of news articles.

3. Gender Bias and Hate Speech Detection

The structure of the BERT-based model for hate speech detection is shown in Fig. 1. Hate speech detection is performed using the BERT embedding vector of the CLS token, which is given to the input vector as a special classification token of the fully connected (FC) layers. Embedding vectors of gender bias or hate speech comments are given by a pretrained BERT model, and the FC layers are tested for one and two layers. In the case of two layers, a hidden layer is set to produce a 1024-dimensional vector, and dropout regulation is applied to the corresponding hidden layer.

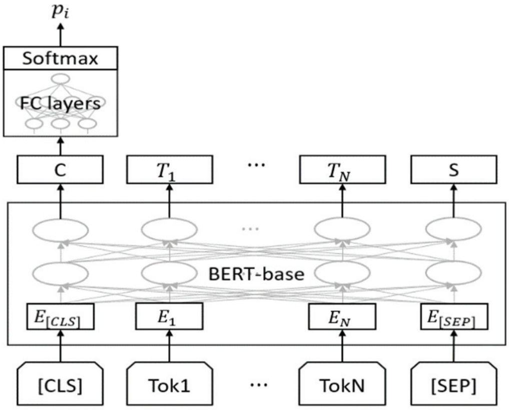


Fig. 1. BERT-based model architecture.

Furthermore, changing the number of FC layers, we attempted to improve performance using ensemble strategies of several models. The structure of the proposed ensemble model for hate speech detection is shown in Fig. 2. First, our ensemble model retrieves predictions $p_1, p_2, \dots, p_k$ of a given sentence x_i from pretrained models $m_1, m_2, \dots, m_k$ trained through transfer learning.

Next, our model applies the logits ensemble to the retrieved predictions p_i, with scaler values λ_i indicating weights. This exploits the characteristics of the pretrained models: as the pretrained models were trained using their corpus, the interpretation of new data might be different. Therefore, even if training is performed using the same dataset, the performance of the models may not be the same. When

$p_1, p_2, \cdots, p_k$ are given as prediction results from k different classifiers, a logits ensemble calculates the final prediction result, as expressed in Eq. (1):

$$p = \lambda_1 p_1 + \lambda_2 p_2 + \cdots + \lambda_k p_k, \tag{1}$$

where $\lambda_i \in \{0, 0.01,\ 0.02, \ldots, 0.99, 1\}$ denotes a set of the best-performing values on the validation set that satisfy $\sum_i \lambda_i\ =\ 1$, as obtained by a depth-first searching algorithm.

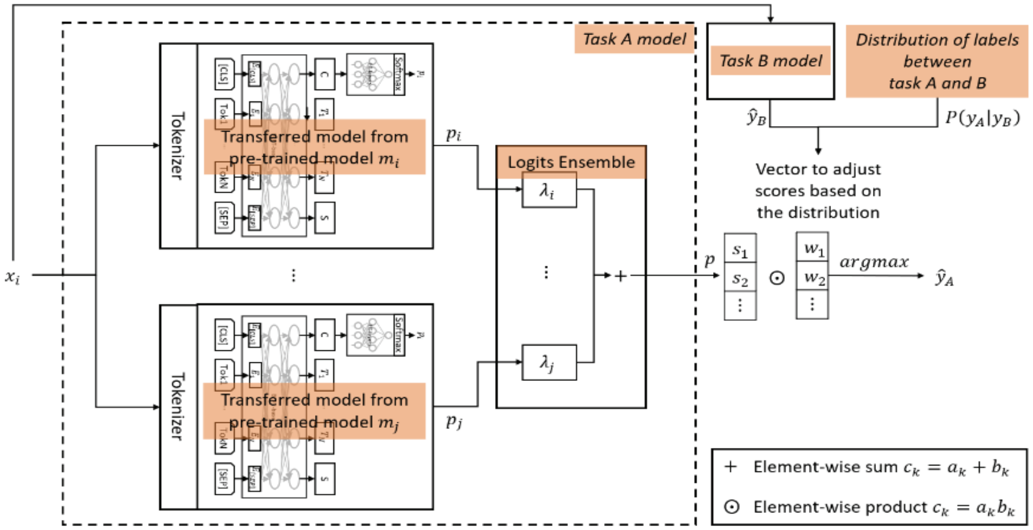

Task A model
Task B model
Distribution of labels between task A and B
$\hat{y}_B$
$P(y_A|y_B)$
Vector to adjust scores based on the distribution
x_i
Tokenizer
[CLS]
Tok1
TokN
[SEP]
C
Softmax
Transferred model from pre-trained model m_i
p_i
Logits Ensemble
λ_i
λ_j
Tokenizer
Transferred model from pre-trained model m_j
p_j
+
p
s_1
s_2
⊙
w_1
w_2
$argmax$
$\hat{y}_A$
+ Element-wise sum $c_k = a_k + b_k$
⊙ Element-wise product $c_k = a_k b_k$

Fig. 2. Logits ensemble with label distribution.

Finally, the model performs an ensemble based on the distribution of labels based on the logits ensemble result p. This ensemble technique uses the statistical correlation between two different tasks A and B that use the same dataset with two hypotheses: 1) the test set must follow the distribution between the two tasks on the validation or training set; 2) the task B model should outperform the task A model. We define the distribution $P(y_A|y_B)$ as the number of data points with the label y_A in task A and the label y_B in task B divided by the number of data points with the label y_B in task B. The technique uses a high-performance trust model for task B and assigns weight to each prediction value $s_i \in p$ for the label A_i in task A according to the distribution $P(y_A\ =\ A_i|y_B\ =\ \hat{y}_B)$, where $\hat{y}_B$ denotes the prediction of the trust model. If $P(y_A|y_B)$ is considered too low, the ensemble model determines whether the prediction of the task A model is likely wrong and allocates a low weight to the prediction value.

Gender bias and hate speech detection is a classification problem with different labels but the same dataset. Therefore, there is a statistical correlation between the different labels. Using this correlation, an ensemble model that performs different classifications can be developed. Our ensemble model was constructed using the label distribution shown in Table 2. From the label distribution, an ensemble between bias detection and hate speech detection can be differentiated. For example, if the result in the bias detection model is not “none,” we lower the weight of the “none” prediction in hate speech detection. In other words, the weighting of the ensemble is performed conditionally using the bias detection model, which has better performance than the hate speech detection model.

Table 2. Distribution of labels in the annotated corpus (unit: %)

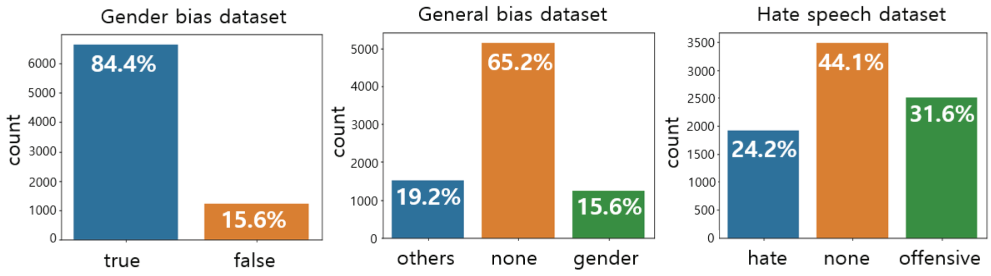
	Hate	Offensive	None	Sum (Bias)
Gender	10.15	0.58	0.98	15.71
Others	7.48	8.94	1.74	18.16
None	7.48	19.13	39.08	65.70
Sum (Hate)	25.11	32.66	41.80	100

For each pretrained model m_i, we not only performed finetuning operations such as changing the number of FC layers but also improved model performance using various other strategies. The main strategies for improving the performance are as follows.

Extreme gradient boosting (XGBoost) is an ensemble technique that makes strong predictions by combining several weak decision trees. Because it cannot be directly applied to natural language, we attempted to improve performance by training the BERT-based models with a multilayer perceptron head and learning XGBoost with the last hidden state output value of the BERT models.

The hate speech dataset is labeled on three types of tags: general bias, gender bias, and hate speech tags. It consists of a train set, a dev set, and a test set that is freely available at https://github.com/kocohub/korean-hate-speech. A total of 9,381 human-labeled comments are split into training, validation, and test sets with 7,896, 471, and 974 test data points, respectively. The test set is unlabeled, but the train and dev sets are labeled with the three tags, as shown in Fig. 3.

The training set has an uneven distribution of observations, which is not a balanced dataset. Because an unbalanced dataset may cause poor performance, we adopted a data sampling strategy to make the dataset more balanced [23]. Data sampling strategies for solving the unbalanced dataset problem include undersampling and oversampling. Undersampling has the disadvantage of significantly reducing the number of data points used in model training. In a situation, where the size of the training set is small, reducing the number of data points degrades model performance. Therefore, we adopted oversampling in which there is no risk of reducing the size of the data. Our dataset was constructed through random shuffling by simply amplifying the data of a smaller size by *n* times.



(a) (b) (c)

Fig. 3. Distribution of class labels for each dataset: (a) gender bias, (b) general bias, and (c) hate speech.

4. Experiments and Results

We experimented with all three tasks, which are gender bias, general bias, and hate speech detection. In each task, hyperparameter tuning was performed to maximize the performance of our model. The aforementioned strategies were used to further improve our model. The experimental environment is

Windows 10 with Intel Core i7-7700 CPU @3.60 GHz, 64 GB RAM, and RTX 3060 GPU. The models were implemented using the PyTorch framework.

4.1 Gender Bias Detection

For each pretraining model, transfer learning for the Korean gender bias detection problem was performed without finetuning. The adjusted hyperparameters are the number of FC layers (1, 2), dropout rate (0.0, 0.1, 0.3, 0.5), learning rate (1e-5, 2e-5, 5e-5), and batch size (16, 32, 64). We used AdamW as an optimizer for training, warmup and linear as a scheduler, and cross-entropy loss as a loss function. Table 3 shows the best performance results for each pretraining model before tuning. The model that showed the best performance on the validation set was the RoBERTa-based model of KLUE, with an F1-score of 0.6145 on the test set, which is lower than the baseline of 0.6814. Accordingly, fine-tuning was performed on the pretrained model. Due to the large size of the BERT model, the maximum batch size was set to 16, as presented in Table 4. During fine-tuning, the learning rate was reduced to prevent the weight of the base model from being significantly modified.

The Soongsil-BERT embedding model exhibited the best performance on the validation set and obtained an F1-score of 0.7416 on the test set. For higher performance, a logits ensemble strategy was used, and the results are shown in Table 5. In Korean gender bias detection, an ensemble of the Soongsil-BERT and KcELECTRA models obtained the highest F1-score of 0.7711.

Table 3. Hyperparameters settings

Base model	FC layers	Dropout rate	Learning rate	Batch size	Best F1-score (validation set)
$\boldsymbol{BERT_{KoBERT_{SKT}}}$	2	0.3	2e-5	16	0.3999
$\boldsymbol{RoBERTa_{soongsil-bert}}$	2	0.1	2e-5	16	0.6363
$\boldsymbol{RoBERTa_{KLUE}}$	2	0.3	5e-5	64	0.6428

Table 4. Hyperparameter settings with fine-tuning

Base model	FC layers	Dropout rate	Learning rate	Batch size	Best F1-score
$\boldsymbol{RoBERTa_{soongsil-bert}}$	2	0.1	1e-5	16	0.8085
$\boldsymbol{RoBERTa_{KLUE}}$	2	0.3	1e-5	16	0.7878
$\boldsymbol{ELECTRA_{KcELECTRA}}$	2	0.5	1e-5	16	0.7812

Table 5. Logits ensemble result for gender bias detection

Ensemble model	Best F1-score		Baseline
	Validation set	Test set	Test set
$\boldsymbol{RoBERTa_{soongsil-bert}} + \boldsymbol{RoBERTa_{KLUE}}$	0.8142	0.7458	0.6814
$\boldsymbol{RoBERTa_{soongsil-bert}} + \boldsymbol{ELECTRA_{KcELECTRA}}$	0.8217	0.7711	

4.2 General Bias Detection

We explored Korean general bias detection. Hyperparameter tuning was performed in a manner similar to Korean gender bias detection. The results are shown in Table 6. In the finetuned layers column of the table, FULL indicates that all layers of the base model are fine-tuned, and 0 indicates that no layers of the base model are fine-tuned. The ELECTRA embedding model exhibited the best performance on the validation set and achieved an F1-score of 0.6955 on the test set, which is higher than the baseline of 0.6326 shown in Table 7. For higher performance improvement, we applied the logits ensemble strategy. The results are shown in Table 7.

An ensemble of the Soongsil-BERT and KcELECTRA models achieved a higher F1-score of 0.7139 on the test set than when using a single model. Because the general bias problem entails a gender bias problem well-trained in Korean gender bias detection, the ensemble method exhibited the best performance in gender bias detection. A higher F1-score of 0.7166 was achieved on the test set.

Table 6. Fine-tuning for general bias detection

Base model	Fine-tuned layers	FC layers	Dropout rate	Learning rate	Batch size	Best F1-score
$\boldsymbol{RoBERTa_{soongsil-bert}}$	0	2	0.1	2e-5	32	0.6597
$\boldsymbol{RoBERTa_{soongsil-bert}}$	FULL	2	0.1	2e-5	16	0.7656
$\boldsymbol{ELECTRA_{KcELECTRA}}$	0	2	0.3	5e-5	32	0.4603
$\boldsymbol{ELECTRA_{KcELECTRA}}$	FULL	2	0.5	1e-5	16	0.7754

Table 7. Logits ensemble result for general bias detection

Ensemble model	Best F1-score		Baseline
	Validation set	Test set	Test set
$\boldsymbol{RoBERTa_{soongsil-bert} + ELECTRA_{KcELECTRA}}$	0.9603	0.7139	0.6326
$\boldsymbol{RoBERTa_{soongsil-bert} + ELECTRA_{KcELECTRA}}$ $\boldsymbol{+ GenderBiasDetectionModel}$	-	0.7166	

4.3 Hate Speech Detection

We applied our model to the hate speech detection task, and the results are shown in Table 8. The KcELECTRA embedding model achieved the highest F1-score on the validation set and an F1-score of 0.6431 on the test set, which is higher than the baseline of 0.5255. An ensemble strategy has been investigated for improvement. We attempted to improve performance using a logits ensemble of the RoBERTa-based model of KLUE and the KcELECTRA model. The ensemble achieved higher performance than the single model, with an F1-score of 0.6524 on the test set. Next, we investigated the best-performing classifier and ensemble in the general bias detection task. Note that no classification class in general bias detection matches a classification class in hate speech detection. Therefore, an ensemble using the label distribution was developed, which achieved an F1-score of 0.6574, which is higher than that of the logits ensemble model.

Table 8. Logits ensemble result for hate speech detection

Ensemble model	Best F1-score		Baseline
	Validation set	Test set	Test set
$\boldsymbol{ELECTRA_{KcELECTRA} + XGBoost}$	0.7442	0.6032	0.5255
$\boldsymbol{RoBERTa_{KLUE} + ELECTRA_{KcELECTRA}}$	0.7435	0.6524	
$\boldsymbol{RoBERTa_{KLUE} + ELECTRA_{KcELECTRA}}$ $\boldsymbol{+ KoreanBiasDetectionModel}$	-	0.6574	

Table 9 shows the rankings of our models in the Kaggle leaderboard (http://www.kaggle.com/c/korean-bias-detection) as of August 20, 2022. We achieved 1st place in both the general bias and gender bias tasks. We attempted to improve our model through additional research on the ensemble strategy using a distribution between general bias and hate speech, achieving the highest performance.

Table 9. Final results on the Kaggle leaderboard

Kaggle competition	Best model	Teams	Ranking	F1-score
Gender bias	$RoBERTa_{soongsil-bert} + ELECTRA_{KcELECTRA}$	17	1	0.7711
General bias	$RoBERTa_{soongsil-bert} + ELECTRA_{KcELECTRA} + GeneralBiasModel$	14	1	0.7166
Hate speech	$RoBERTa_{KLUE} + ELECTRA_{KcELECTRA} + GeneralBiasModel$	65	4	0.6574

5. Discussion

Our quantitative evaluation has demonstrated the effectiveness of the proposed ensemble approach. We showed significant performance improvement using our proposed logits ensemble by achieving 10.8%, 2.64%, and 1.43% performance improvement for gender bias, general bias, and hate speech tasks, respectively. The logits ensemble outperforms the other models on binary classification problems, improving performance by approximately 4–7 times. Moreover, we improved performance by proposing a distribution-based ensemble technique that achieved 0.38% and 0.77% higher performance on the general bias and hate speech tasks, respectively. Because we performed this technique only for cases with significantly low distributions, it achieved a lower improvement than the logits ensemble. We believe that this technique can exhibit significant improvement on tasks that have many classes, where distribution values $P(y_A|y_B)$ are more varied.

Zimmerman et al. [21] developed a logits ensemble method with equal weights to the result vectors of a single model and achieved an improvement of approximately 2%. However, our logits ensemble model based on a weighting method exhibited higher performance with an improvement of 1.43%–10.8%. Moreover, in contrast to our model, which used only 3–4 pretrained models, they used 10 models to achieve such improvement. This indicates the proposed logits ensemble method that uses a depth-first searching algorithm can fully utilize the capability of ensembled models and achieve high performance even with a few pretrained models.

Karim et al. [22] proposed a voting-based ensemble technique using various pretrained models and achieved approximately 1.15% performance improvement. However, our ensemble model with a label distribution achieved a higher improvement of 1.44%–2.65%. This demonstrates the strong competitiveness of our proposed ensemble model compared with widely used ensemble techniques such as voting and boosting.

Limits of our approach and potential improvement directions:

Despite the good performance improvement of the proposed ensemble model, several directions could still be explored. First, our proposed logits ensemble model uses depth-first searching to find weights λ_i.

This can be a critical issue in terms of speed, when the number of classes is increased or if we want to finetune weights. Various optimization methods, such as Bayesian optimization or even neural network models, can be used to better tune weights. Second, in the distribution-based ensemble, we adjusted the weights manually. We can improve performance by finding a way to compute effective weights from a given label distribution.

6. Conclusion

Cyberbullying and malicious comments by anonymous users cause severe social problems in cyberspace. It has resulted in the suicide of a celebrity. To solve this problem, we explored the three detection tasks of general bias, gender bias, and hate speech. We proposed a logits ensemble model that combines various models using a weighting scheme. Our ensemble model achieved higher performance than single models. Moreover, we achieved the best performance by developing an ensemble with a label distribution of general and gender bias datasets, where there is a statistical correlation between the labels. Zimmerman et al. [21] developed a logits ensemble method with equal weights and achieved an improvement of approximately 2%. However, our logits ensemble model based on a weighting method and label distribution exhibited better performance with an improvement of 1.43%–10.8%. Malicious comment and hateful expression detection has been investigated; however, it has limitations because there are many variants of hate expressions. Our ensemble model improved the accuracy of bias and hate speech detection. As further research, we are constructing a large-scale dataset for the toxic and hate speech domain.

Acknowledgement

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1F1A1061433) and partly by the Synapsoft ATC+ grant funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

References

[1] A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing,” in *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, 2017, pp. 1-10. https://doi.org/10.18653/v1/w17-1101

[2] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, “Content-driven detection of cyberbullying on the Instagram social network,” in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, New York, NY, 2016, pp. 3952-3958.

[3] J. H. Park and P. Fung, “One-step and two-step classification for abusive language detection on twitter,” in *Proceedings of the 1st Workshop on Abusive Language Online*, Vancouver, Canada, 2017, pp. 41-45. https://doi.org/10.18653/v1/w17-3006

[4] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, “L-HSAB: a Levantine twitter dataset for hate speech and abusive language,” in *Proceedings of the 3rd Workshop on Abusive Language Online*, Florence, Italy, 2019, pp. 111-118. https://doi.org/10.18653/v1/W19-3512

[5] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, “Men also like shopping: reducing gender bias amplification using corpus-level constraints,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017, pp. 2979-2989. https://doi.org/10.18653/v1/d17-1323

[6] S. Kiritchenko and S. M. Mohammad, “Examining gender and race bias in two hundred sentiment analysis systems,” in *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, New Orleans, LA, 2018, pp. 43-53. https://doi.org/10.18653/v1/s18-2005

[7] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, “Gender bias in neural natural language processing,” *Logic, Language, and Security*. Cham, Switzerland: Springer, 2020, pp. 189-202. https://doi.org/10.1007/978-3-030-62077-6_14

[8] Z. Ahmed, B. Vidgen, and S. A. Hale, “Tackling racial bias in automated online hate detection: towards fair and accurate detection of hateful users with geometric deep learning,” *EPJ Data Science*, vol. 11, article no. 8, 2022. https://doi.org/10.1140/epjds/s13688-022-00319-9

[9] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content,” in *Proceedings of the 25th International Conference on World Wide Web*, Geneva, Switzerland, 2016, pp. 145-153. https://doi.org/10.1145/2872427.2883062

[10] H. Rizwan, M. H. Shakeel, and A. Karim, “Hate-speech and offensive language detection in roman Urdu,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Virtual Event, 2020, pp. 2512-2522. http://dx.doi.org/10.18653/v1/2020.emnlp-main.197

[11] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, “Emotionally informed hate speech detection: a multi-target perspective,” *Cognitive Computation*, vol. 14, pp. 322-352, 2022. https://doi.org/10.1007/s12559-021-09862-5

[12] P. Fortuna, J. Soler-Company, and L. Wanner, “How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?,” *Information Processing & Management*, vol. 58, no. 3, article no. 102524, 2021. https://doi.org/10.1016/j.ipm.2021.102524

[13] N. S. Mullah and W. M. N. W. Zainon, “Advances in machine learning algorithms for hate speech detection in social media: a review,” *IEEE Access*, vol. 9, pp. 88364-88376, 2021. https://doi.org/10.1109/ACCESS.2021.3089515

[14] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,’ in *Proceedings of the 26th International Conference on World Wide Web Companion*, Perth, Australia, 2017, pp. 759-760. https://doi.org/10.1145/3041021.3054223

[15] R. Alshalan and H. Al-Khalifa, “A deep learning approach for automatic hate speech detection in the saudi twittersphere,” *Applied Sciences*, vol. 10, no. 23, article no. 8614, 2020. https://doi.org/10.3390/app10238614

[16] O. Sharif and M. M. Hoque, “Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers,” *Neurocomputing*, vol. 490, pp. 462-481, 2022. https://doi.org/10.1016/j.neucom.2021.12.022

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” 2018 [Online]. Available: https://arxiv.org/abs/1810.04805.

[18] J. Moon, W. I. Cho, and J. Lee, “BEEP! Korean corpus of online news comments for toxic speech detection,” in *Proceedings of the 8th International Workshop on Natural Language Processing for Social Media*, Virtual Event, 2020, pp. 25-31. https://doi.org/10.18653/v1/2020.socialnlp-1.4

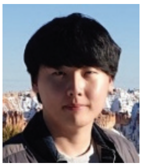
[19] W. I. Cho, J. W. Kim, S. M. Kim, and N. S. Kim, “On measuring gender bias in translation of gender-neutral pronouns,” in *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, Florence, Italy, 2019, pp. 173-181. https://doi.org/10.18653/v1/W19-3824

[20] W. I. Cho and J. Moon, “A study on the construction of Korean Hate speech corpus: based on the attributes of online toxic comments,” in *Proceedings of Annual Conference on Human and Language Technology*, Virtual Event, 2020, pp. 298-303.

[21] S. Zimmerman, U. Kruschwitz, and C. Fox, “Improving hate speech detection with deep learning ensembles,” in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, 2018, pp. 2546-2553.

[22] M. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, M. A. Hossain, and S. Decker, “DeepHateExplainer: explainable hate speech detection in under-resourced Bengali language,” in *Proceedings of 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, Porto, Portugal, 2021, pp. 1-10. https://doi.org/10.1109/DSAA53316.2021.9564230

[23] K. Webster, M. Recasens, V. Axelrod, and J. Baldridge, “Mind the GAP: a balanced corpus of gendered ambiguous pronouns,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 605-617, 2018. https://doi.org/10.1162/tacl_a_00240



Sanggeon Yun **https://orcid.org/0000-0002-0488-9666**

He received his B.S. degree in Computer Science in 2023 at Kookmin University, Seoul, South Korea. In 2023 he started his Ph.D. research at the Bio-Inspired Architecture and Systems (BIASLab) of the University of California, Irvine with Professor Mohsen Imani. His research interests include hyperdimensional computing, machine learning, natural language processing, human-computer interaction, and information visualization.

Seungshik Kang **https://orcid.org/0000-0003-3318-6326**

He received B.S. degree in Computer Science from Seoul National University in 1986, and M.S. and Ph.D. degree in Computer Science from the same University, in 1988 and 1993, respectively. Currently, he is working for Kookmin University as a full professor. His research interests include natural language processing, information retrieval, text mining, big data processing, and machine learning.

Hyeokman Kim **https://orcid.org/0000-0001-7129-6759**

He received his B.S., M.S., and Ph.D. degrees in computer science and engineering from Seoul National University, Seoul, Korea in 1985, 1987 and 1996 respectively. From 1996 to 1999, he was engaged in research on digital video library for the Multimedia Technology Research Laboratory, Korea Telecom where he was a senior member of technical staff. Currently, he is working for Kookmin University as a full professor. His research interests include database, big data processing and modeling.