Federated Learning and LLM-based Social Media Comment Classification System Using Crowdsourcing Techniques

Jungho Kang†
kjh7548@naver.com
† Baewha Women's University, Seoul, 03039 Korea

Summary
Currently, on social media, malicious comments have emerged as a serious issue. Existing artificial intelligence-based comment classification systems have limitations due to data bias and overfitting. To address this, this study proposed a novel comment classification system that combines crowdsourcing and federated learning. This system collects data from various users and utilizes a large language model like KoBERT through federated learning to classify comments accurately while protecting user privacy. It is expected to provide higher accuracy than existing methods and improve significantly the efficiency of detecting malicious comments. The proposed system can be applied to social media platforms and online communities.

Keywords:
SNS, Comment Classification, Crowdsourcing, LLM-based KoBERT, Federated Learning

1. Introduction

With the proliferation of the Internet and social networking services (SNS), comments have become an essential part of online communities, facilitating information sharing and dialogue. However, the anonymity and immediacy of comments make it easy for users to express emotional reactions that can sometimes lead to conflict. Moreover, although comments provide various perspectives on social issues, the anonymity and unchecked freedom of expression inherent in comments can be abused, leading to problems such as negative or aggressive malicious comments [1].

Malicious comments appear in forms such as personal attacks, hate speech, ridicule, and spread of false information, with serious impacts on individuals and society. The seriousness of malicious comments has been highlighted by numerous cases where prominent celebrities have endured severe emotional distress, with some unfortunately resorting to taking their own lives. Furthermore, this issue has gone beyond the entertainment industry—affecting sectors like politics and sports—and has become a widespread societal concern. Existing systems designed to detect artificial intelligence (AI)-based malicious comments have been developed to address these issues but face limitations such as overfitting caused by data bias [2].

Manuscript received October 5, 2024
Manuscript revised October 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.10.3

To address these issues, this paper proposes a social media comment classification system based on federated learning and large language models (LLM) using crowdsourcing techniques. This system is called Federated Learning and LLM-based Comment Classification Algorithm with Crowdsourcing (FLCAC). By collecting large-scale data from various users through crowdsourcing, it is possible to reduce data bias and improve the accuracy of model training. Federated learning techniques offer the advantage of training models simultaneously on multiple devices while protecting the privacy of individual data. Additionally, the FLCAC-based comment classification system can more accurately understand the context and meaning of comments, enabling efficient detection of malicious comments, improving the accuracy of detection, and ultimately fostering a positive SNS communication environment.

2. Related research

Recent studies have focused on significantly improving comment classification performance by using Korean-specific models like Korean bidirectional encoder representations from transformers (KoBERT) and LLM-based KoBERT models. These models demonstrate strengths of overcoming limitations in Korean natural language processing (NLP) tasks and effectively classifying comments by deeply understanding the context and meaning of the text. Additionally, with the combination of technologies like crowdsourcing and federated learning, various approaches are being proposed to implement more efficient, accurate comment classification systems. This study sought to explore the development directions of a comment classification system using these advanced technologies and to propose ways of improving user experience and enhancing the quality of online communities.

2.1 KoBERT

KoBERT is a BERT-based model developed by SKTBrain to enhance the performance of Korean NLP tasks. As a pre-trained model developed by Google, BERT

exhibits high performance in various NLP tasks after being trained on large-scale foreign language datasets [3]. However, it has limitations in terms of accuracy when processing the Korean language.

To overcome these limitations, KoBERT was designed as a model optimized for Korean by learning from a corpus consisting of millions of Korean sentences [4]. Trained on Korean data collected from various sources such as Wikipedia and news articles, KoBERT has demonstrated improved performance in various NLP applications including text classification, sentiment analysis, and question-answering systems. Due to these characteristics, KoBERT has established itself as a crucial tool in Korean NLP research and practical applications, contributing to the advancement of Korean-based AI technologies.

Muhammad Fikri Hasani et al. [5] investigated the combination of preprocessing techniques and text vectorization in investor comment classification using the KoBERT model. Their study proposed a method of enhancing the performance of machine learning models by combining various preprocessing methods and vectorization techniques, thereby verifying the effective utilization of KoBERT.

2.2 LLM-based KoBERT

LLMs demonstrate excellent performance in various NLP tasks, and they are effectively used in comment classification systems. According to recent research, LLMs have a strong ability to understand deeply the context and meaning of the text, making them effective in classifying comments as positive, negative, or neutral [6].

Huanhuan Zhao et al. [7] significantly improved text classification performance by augmenting existing datasets with data generated by LLMs such as ChatGPT. Their study compared two methods: generating new data and rewriting the existing dataset. They generally found that generating new data resulted in better performance.

Mayank Mishra et al. [8] demonstrated that using pseudo-code prompts with an LLM-based KoBERT model improved performance in comment classification tasks. Their study showed that the use of pseudo-code prompts led to a 7~16-point increase in F1 scores and a 12~38% improvement in ROUGE-L scores, validating the approach's effectiveness in comment classification tasks.

2.3 Crowdsourcing

As a compound word combining "crowd" and "outsourcing," crowdsourcing refers to a method of solving problems by making specific tasks available to the public and encouraging voluntary participation [9]. A prime example is Wikipedia, an online encyclopedia where users create and edit content directly; thus maintaining the accuracy and diversity of information through collective intelligence [10].

Hanwei Zhu et al. [11] proposed a blockchain-based decentralized, anonymous crowdsourcing mechanism, an open anonymous participation system that encourages the involvement of users without revealing their identity; thus ensuring data privacy and promoting the active participation of a wide range of users. Additionally, L. Ang et al. [12] proposed a new approach that combines Internet of Things (IoT) technology with crowdsourcing for use in information collection and processing.

Hanna Abi Akl et al. [13] combined machine learning and LLM to improve the performance of code comment classification. Conducted as part of the FIRE 2023 shared task, their study utilized crowdsourcing for comment classification and validated that the adoption of LLMs resulted in improved performance compared to traditional methods.

These previous studies demonstrate that crowdsourcing can be successfully applied in various technical environments and suggest that it can be effectively utilized for specific tasks such as comment classification.

2.4 Federated learning

Federated learning (FL) is a decentralized learning approach where multiple participants collaboratively train a machine learning model using local data without centralized data storage. In this method, data remain on the participants' devices, and only model parameters are transmitted; thus maintaining data privacy while enabling collaborative learning [14]. FL can be applied in various environments while maintaining data security, making it useful in fields where privacy is crucial such as healthcare, finance, and smart cities.

Shubham et al. [15] proposed a system that integrates blockchain and FL in IoT environments to ensure data storage security and address issues related to data leak and privacy protection. This system enhances security among vehicles, roadside devices, and base stations by preventing data tampering and securely sharing distributed models.

Nair et al. [16] proposed Fed_select, a privacy-preserving FL framework that combines edge computing and differential privacy for big data analysis in the Internet of Medical Things (IoMT) environments. This framework enhances efficiency through real-time analysis and high-speed processing while ensuring user anonymity and minimizing the risk of data leak.

2.5 Previous studies on comment classification systems

To improve the social media comment environment, previous studies have focused on effectively managing user-generated content through sentiment analysis, topic classification, and identification of spam and malicious comments. In recent years, focus has been on developing systems that utilize deep learning-based language models like BERT to achieve higher performance in understanding context and distinguishing semantic nuances. The following studies have been conducted in this area:

P. Juyal et al. [17] conducted sentiment analysis on text using a hybrid deep learning technique that combines convolutional neural networks (CNN) and support vector machines (SVM). Their study focused on enhancing the efficiency of sentiment classification by integrating the unique feature extraction capabilities of CNNs into the superior classification performance of SVMs. The CNN-SVM combination demonstrated higher accuracy and generalization performance in sentiment analysis compared to single models and performed well even with noisy data.

Prashant Giridhar Shambharkar et al. [18] evaluated the performance of various deep learning models including BiLSTM and CNN for identifying harmful comments such as cyberbullying and hate speech. They analyzed how each model handles complex emotional expressions. BiLSTM excels in understanding emotions in long contexts by considering temporal dependencies, whereas CNN performs well in quickly identifying important keywords in the text. The research contributed to designing a more efficient sentiment analysis model for creating a safer online community environment by integrating the strengths and weaknesses of each model.

Xinyue Feng et al. [19] proposed an integrated algorithm for sentiment analysis of data from Chinese SNS Weibo. This algorithm combines the capability of CNNs to capture the local features of text quickly with the ability of BiLSTM to retain long-context information, significantly enhancing the accuracy of text sentiment analysis. Additionally, by incorporating an attention mechanism, the model focuses on key text features to optimize analysis efficiency. This approach achieved higher accuracy compared to existing benchmark models and proved particularly useful for real-time sentiment analysis applications.

3. Proposed system

To improve SNS environments, this paper proposes a comment classification system based on federated learning and LLMs using crowdsourcing techniques called FLCAC. Figure 1 shows the system proposed in this paper.

The FLCAC-based social media comment classification

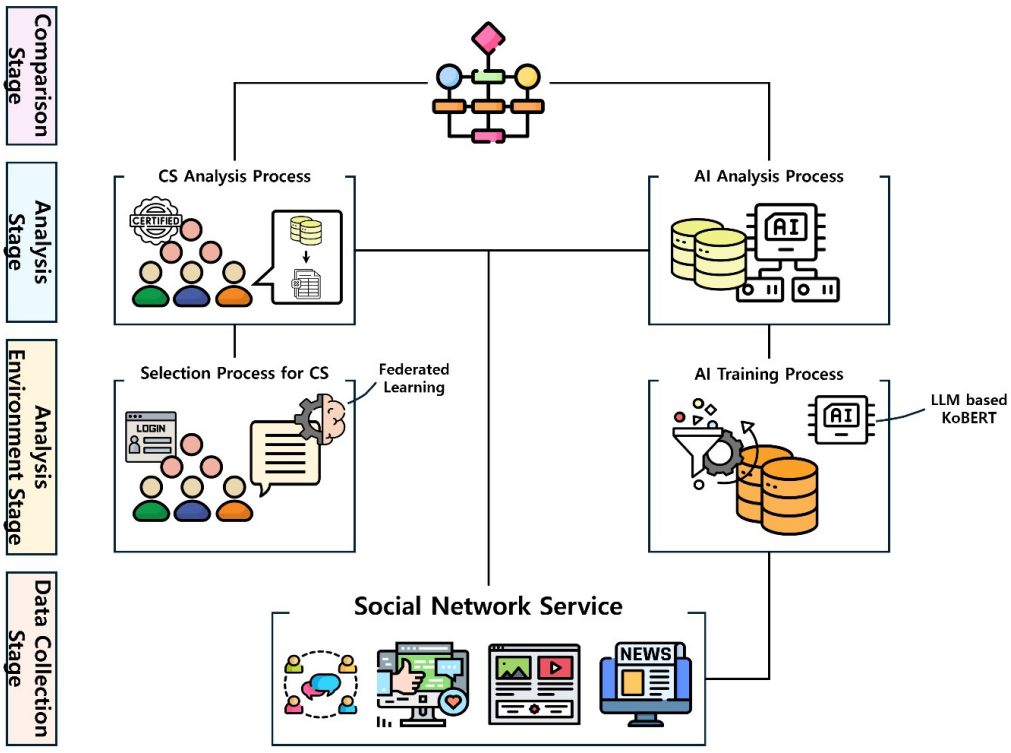

Fig 1. FLCAC-based comment classification system scenario

system performs comment classification through a total of five stages as follows:

- Data collection stage: Collect various comment data in the social media environment for use as material for learning and analysis.
- Analysis environment setup stage: The analysis environment setup stage is divided into two parts. During this stage, the process of selecting crowds for crowdsourcing is implemented to prevent malicious activities. Additionally, a portion of the collected data is used for AI training to optimize the analysis of social media comments.
- Analysis stage: Analyze the collected social media comment data using both crowdsourcing and trained AI model.
- Comparison stage: Compare the results from the crowdsourcing and the trained AI model at a 60:40 ratio to derive the final classification outcome for the comments.

3.1 System scenarios

3.1.1 Data collection and analysis environment setup stage

In the data collection stage, comment crawling is used to gather various comment data from the SNS environment. The collected data are then divided into training data and additional collected data. Subsequently, a crowdsourcing and AI environment for comment analysis is set up.

The crowd selection process in the analysis environment setup stage is designed to prevent malicious crowdsourcing. This process can be reviewed through the crowd selection procedure for crowdsourcing, as illustrated in Figure 2.

In the case of crowdsourcing, participation with malicious intent can result in lower accuracy compared to existing AI training models. To prevent this, user authentication and verification processes are essential. In this process, the SNS login information of the user is used to verify his/her identity, and sentiment analysis of his/her past comments is conducted to determine whether he/she is a malicious user.

In particular, since comments often contain sensitive personal information, FL is employed to protect user privacy. Sentiment analysis is performed on the user's personal device, ensuring the privacy of personal information. Only the calculated sentiment analysis parameters are transmitted to the cloud, which then examines these parameters to select users suitable for crowdsourcing. This process enhances the reliability of crowdsourcing and protects the system from malicious participation.

In the AI learning process of the analysis environment setup stage, a system that classifies social media comments into positive, neutral, and negative is built using the KoBERT model. To achieve this, comments are randomly collected through crawling from various social media platforms such as YouTube and Instagram during the data collection stage. The collected data are divided into training data for the KoBERT model and test data for evaluating the model's accuracy. The ratio of training data to test data is 9:1, which is commonly used when comparing the overfitting error rate of AI models [20].

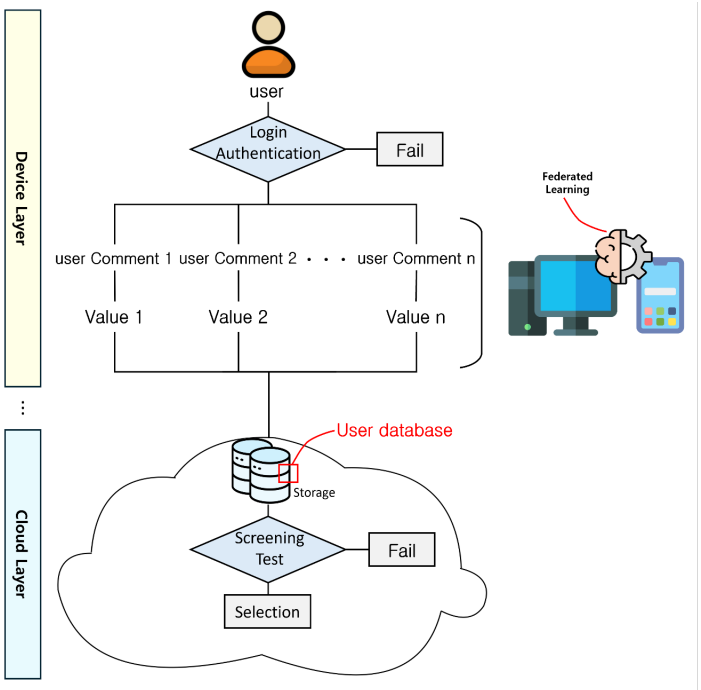

Fig 2. Crowd selection flowchart for crowdsourcing

Algorithm 1: KoBERT Model Training

```
Input: training data batches, learning rate scheduler, cross-entropy loss function
Output: training accuracy
Process:
Begin
1.    Initialize training accuracy and test accuracy to 0.0
2.    For each epoch e in range num_epochs:
3.        #Training Phase
4.        Set 'model' to training mode
5.        For each batch (token_ids, valid_length, segment_ids, label) in train_dataloader:
6.            Reset gradients: optimizer.zero_grad()
7.            Move token_ids, segment_ids, and label to the computing device
8.            Compute model output: out = model(token_ids, valid_length, segment_ids)
9.            Calculate loss: loss = loss_fn(out, label)
10.           Perform backpropagation: loss.backward()
11.           Apply gradient clipping:
12.               torch.nn.utils.clip_grad_norm_(model.parameters(), max_grad_norm)
13.           Update model parameters: optimizer.step()
14.           Adjust learning rate: scheduler.step()
15.           Update training accuracy: train_acc += calc_accuracy(out, label)
16.           If batch_id is a multiple of log_interval
17.               then log the current epoch, batch ID, loss, and training accuracy
18.       Log total training accuracy for epoch e:
19.               print("epoch {} train acc {}".format(e+1, train_acc / (batch_id+1)))
20.       #Evaluation Phase
21.       Set model to evaluation mode
22.       For each batch (token_ids, valid_length, segment_ids, label) in test_dataloader:
23.           Move token_ids, segment_ids, and label to the computing device
24.           Compute model output: out = model(token_ids, valid_length, segment_ids)
25.           Update testing accuracy: test_acc += calc_accuracy(out, label)
26.       Log total testing accuracy for epoch e.
27.   End For (epoch loop)
28.   Return total training accuracy and test accuracy
29.end
End
```

Algorithm 1. KoBERT model training algorithm

The KoBERT model is trained through a learning algorithm composed of a training phase and an evaluation phase, as presented in Algorithm 1. In the training phase, the model learns to understand the meaning of the given comments effectively and predict the matching category accurately based on the training data. The evaluation phase assesses the performance of the trained model using test data, verifying the accuracy and reliability of its predictions.

3.1.2 Analysis and comparison stage

In the analysis stage, the collected social media comment data are analyzed using both crowdsourcing and trained KoBERT model.

In the comparison stage, the final classification of social media comments is based on the analysis results from both crowdsourcing and KoBERT model. If the analysis results are consistent, the corresponding category is determined as the final classification of the comment. If the crowdsourcing and KoBERT model analysis results do not match, however, an additional procedure is conducted to compare the two results and determine the final comment classification. In this process, the final result is derived by applying a 60:40 ratio to the analysis results from crowdsourcing and AI model, respectively. This ratio is based on a previous study [21] that compared various ratios of crowdsourcing and AI model analysis ranging from 9:1 to 1:9 and demonstrated that the 60:40 ratio provided the highest accuracy. Therefore, this study adopted the same ratio to determine the final comment classification.

By combining the collective intelligence of crowdsourcing with the automated analytical capabilities of the AI model, this approach seeks to produce more reliable and accurate comment classification results. This method seeks to maximize the complementary roles of crowdsourcing and AI model, ultimately enhancing the efficiency of social media comment classification.

3.2 Discussion

This paper introduced a novel comment classification method that integrates Federated Learning (FL), LLM, and crowdsourcing techniques to address the limitations of traditional Artificial Intelligence (AI)-based methods. These conventional systems often suffer from data bias and overfitting, which reduce the accuracy and reliability of detecting malicious comments. FLCAC (Federated Learning and LLM-based Comment Classification Algorithm with Crowdsourcing) system offers several advancements that enhance confidentiality, integrity, availability, bias reduction, and accuracy as follows:

- **Confidentiality:** The system leverages federated learning (FL) to ensure that user data remains private.

Data from users is not centrally stored; instead, only model parameters are transmitted, keeping personal data on users' devices. This protects the confidentiality of user information, including sensitive comment data.

- **Data Bias Reduction:** By using crowdsourcing and federated learning, the system minimizes overfitting and biases that could compromise the integrity of the model's outputs. Crowdsourcing ensures that diverse data points from various users are collected, which strengthens the accuracy and fairness of the classification.
- **Integrity:** During the federated learning process, only the parameters are shared, reducing the risk of data manipulation. Furthermore, the system uses crowdsourcing to cross-verify the AI model's output, ensuring reliable and consistent classification results.
- **Availability:** The federated learning approach ensures that the model can be trained across multiple devices, which enhances system availability and scalability. The classification system is designed to work across various platforms, ensuring continuous availability and real-time comment analysis on social media.
- **Accuracy:** The proposed system combines crowdsourcing and federated learning to ensure diverse data collection, which mitigates biased training and enhances the model's prediction accuracy. By aggregating data from a wide range of users, the system effectively reduces overfitting, improving its ability to classify comments more precisely. Leveraging KoBERT, a Korean-specific large language model, the system gains a deep understanding of the context and meaning behind comments, allowing for more accurate detection of malicious content compared to traditional AI-based systems. To further ensure accuracy, the system compares the classification results from both crowdsourcing and the AI model. When the results align, the classification is confirmed; if they diverge, an additional review process is initiated, reinforcing the reliability of the final output.

4. Conclusion

Existing AI-based comment classification systems have seen reduced accuracy due to data bias and overfitting during the training process. To address the issue of malicious comments on social media, this study proposed a comment classification system based on FL and LLMs with crowdsourcing techniques called FLCAC. The proposed system protects the privacy of crowdsourcing users and identifies malicious users through an FL-based user authentication process, thereby preventing malicious crowdsourcing in the analysis process. In addition, it ensures higher accuracy and reliability compared to existing methods by adopting an LLM-based KoBERT model that can more accurately understand the context and meaning of comments.

This study aimed to enhance learning accuracy by collecting large-scale data from diverse users through crowdsourcing—including various opinions—and by detecting malicious comments that may have been omitted due to overfitting. This approach offers a new direction for detecting malicious comments, fosters a reliable SNS communication environment, and contributes to the development of a positive online community culture. Future research will focus on exploring the scalability of this system across various social media platforms and language environments and will include simulations to validate the system's performance and effectiveness.

Acknowledgment

This work was financially supported by Baewha Women's University.

References

[1] Wang, Jyun-Cheng, Gabriel Indra Widi Tamtama, and Retnani Latifah. "Civil or Uncivil: Seeing the Role of User Actors and Providing Comments on Social Media." 2024 21st International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). IEEE, pp.001-006, 2024.

[2] M Behnke, N Briner, D Cullen, K Schwerdtfeger, J Warren, R Basnet, T Doleck, "Feature engineering and machine learning model comparison for malicious activity detection in the dns-over-https protocol.", IEEE Access, 9, 12, pp.129902-129916, 2021.

[3] VG Tarun, R Sivasakthivel, G Ramar, M Rajagopal, G Sivaraman, "Exploring BERT and Bi-LSTM for Toxic Comment Classification: A Comparative Analysis." 2024 Second International Conference on Data Science and Information System (ICDSIS). IEEE, pp.001-006, 2024.

[4] https://github.com/SKTBrain/KoBERT

[5] Hasani, Muhammad Fikri, "Investigating the Combination between Pre-processing Technique and Text Vectorization for Machine Learning Model in Investor Comment Classification." 2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP). IEEE, pp.001-005, 2023.

[6] Niharika Prasanna Kumar, Kishore Srinivasan, Dhanesh Ramesh, "Analyzing Public Sentiment Towards LLM: A Twitter-Based Sentiment Analysis", 2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-RVITM). IEEE, pp.001-008, 2023.

[7] H Zhao, H Chen, TA Ruggles, Y Feng, D Singh, HJ Yoon, "Improving Text Classification with Large Language Model-Based Data Augmentation." Electronics, 13,13, pp.001-014, 2024.

[8] M Mishra, P Kumar, R Bhat, R Murthy V, D Contractor, S Tamilselvam, "Prompting with Pseudo-Code Instructions", The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.001-020, 2023.

[9] Kadija M. Tahlil, Ucheoma Nwaozuru, Donaldson F. Conserve, Ujunwa F. OnyeamaID, Victor OjoID, Suzanne Day, Jason J. Ong, Weiming Tang, Nora E. Rosenberg, Titi Gbajabiamila, Susan Nkengasong, Chisom Obiezu-Umeh, David Oladele, Juliet Iwelunmor, Oliver EzechiID, Joseph D. Tucker, "Crowdsourcing to support training for public health: A scoping review." PLOS global public health, 3, 7, pp.001-017, 2023.

[10] Wonsik Shim, Jeayeon Byun, Minjung Kim, "Analysis of Wikipedia Citations in Peer-Reviewed Journal Articles", Journal of the Korean Society for Library and Information Science, 42, 2, pp.247-264, 2013.

[11] H Zhu, N Wang, SCK Chau, M Khonji, "Blockchain-enabled Decentralized Anonymous Crowdsourcing Based on Anonymous Payments." 2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). IEEE, pp.001-002, 2023.

[12] Ang, Kenneth Li Minn, Jasmine Kah Phooi Seng, Ericmoore Ngharamike, "Towards crowdsourcing internet of things (crowd-iot): Architectures, security and applications." Future Internet, 14, 2, pp.001-050, 2022.

[13] Hanna Abi Akl, "A ML-LLM pairing for better code comment classification", Forum for Information Retrieval Evaluation 2023, pp001-011, 2023.

[14] Hao Li, Chengcheng Li, Jian Wang, Aimin Yang, Zezhong Ma, Zunqian Zhang, Dianbo Hua, "Review on security of federated learning and its application in healthcare", Future Generation Computer Systems, 144, pp.271-290, 2023.

[15] Shubham, Gagandeep, Vidushi Agarwal, Sujata Pal, "IoT Data Security: An Integration of Blockchain and Federated Learning", 2023 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, pp.001-006, 2023.

[16] Nair, Akarsh K., Jayakrushna Sahoo, and Ebin Deni Raj, "Privacy preserving Federated Learning framework for IoMT based big data analysis using edge computing." Computer Standards & Interfaces 86, 103720, pp.001-020, 2023.

[17] Juyal, Prachi, and Amit Kundaliya. "A Comparative Study of Hybrid Deep Sentimental Analysis Learning Techniques with CNN and SVM." 2023 IEEE World Conference on Applied Intelligence and Computing (AIC). IEEE, pp.001-005, 2023.

[18] PG Shambharkar, H Singh, HR Raghav, H Verma, "Exploring the Efficacy of Deep Learning Models for Multiclass Toxic Comment Classification in Social Media Using Natural Language Processing." 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, pp.001-008, 2023.

[19] Feng, Xinyue, Niwat Angkawisittpan, and Xiaoqing Yang, "A CNN-BiLSTM algorithm for Weibo emotion classification with attention mechanism." Mathematical Models in Engineering, 10, 2, pp.001-011, 2024.

[20] Lior Zoref, "Mindsharing: The Art of Crowdsourcing Everything", Penguin Putnam, 2015.

[21] Heeji Park, Jimin Ha, Hyaelim Park, Jungho Kang, "Comment Classification System using Deep Learning Classification Algorithm based on Crowdsourcing", Proceedings of the Korea Information Processing Society Conference, 11a, pp.864-867, 2021.

Jungho Kang received a Ph.D. in Engineering, specializing in Computer Communications, from the Department of Computer Science at Soongsil University in February 2014. Since March 2018, Dr. A has been serving as a faculty member in the Department of Software Engineering at Baewha Women's University. Dr. A is also the Chief Editor of the Journal of Human-centric Computing (HCIS) and the President of the Korea Computer Industry Association (KCIA). In addition, Dr. A actively participates in the organizing committees of numerous international conferences, continuing to contribute to the global academic community.