LLM-Assisted Labeling and Transformer Model Ensemble for Disaster-Related News Classification in Korean Regional Media (2011–2024)

Abstract

Timely situational awareness is critical during disasters, and news media serve as a vital source for rapid and reliable disaster-related information. However, existing news-mining studies primarily focus on a limited set of natural hazards, often neglecting diverse man-made disasters. To address this limitation, we propose a hybrid classification framework combining commercial API-served large language models (C-LLMs) and open-source fine-tunable pre-trained language models (F-PLMs) to classify 31 diverse disaster types. We first automatically labeled 56,563 stratified news headlines from 2.77 million articles using two C-LLMs (GPT-3.5-Turbo and Gemini-2.0 Flash-Lite). Human reviewers manually adjudicated only discrepant cases (19.1%), significantly reducing manual labeling effort. This labeled dataset was then used to fine-tune two transformer-based F-PLMs (BERT and KoELECTRA), which were combined into a soft-voting ensemble. The ensemble achieved a Macro-F1 score of 0.9589, closely approaching the best-performing C-LLM (Gemini, 0.9817). Temporal analysis of classified news articles showed strong alignment with 92 official activation dates of Korea's Central Disaster and Safety Countermeasures Headquarters, validating the framework's real-world applicability. The proposed hybrid approach is currently utilized by Korea's National Disaster Management Research Institute for real-time disaster news monitoring and serves as one of the essential references for governmental disaster-budget prioritization.

Keywords: large language model, pre-trained language model, disaster-related news classification, disaster management

1. Introduction

Rapid and accurate information acquisition is crucial in disaster situations, and news media play a central role in providing such information [1]. News outlets deliver reliable and timely information derived from official governmental announcements and on-site reporting, thus serving as authoritative sources for disaster response [2]. Consequently, news data are widely acknowledged as valuable resources for disaster preparedness and response strategies [3].

Recent studies have utilized news articles to estimate disaster impacts or to supplement existing global disaster databases. For example, De Brito et al. [4] demonstrated the capability of swiftly monitoring drought impacts through text-mining of media coverage during the 2018/19 drought in Germany. Furthermore, regional news outlets often provide detailed reports on small-scale disasters, which are typically overlooked by global disaster databases. Integrating local news data can thus significantly enhance the quality of global disaster impact databases. A case study conducted in Malawi illustrated that local newspaper data effectively supplemented official statistics by filling spatiotemporal gaps in flood impact assessments [5]. Consequently, news data are recognized as valuable supplementary sources for disaster management. While these studies demonstrate the potential of news data, their scope has largely been confined to specific, high-frequency natural hazards such as floods and droughts. A significant research gap remains in the systematic classification of a broader spectrum of disasters, especially the diverse and multifaceted category of man-made disasters, which are often overlooked in large-scale automated analyses.

Advancements in Natural Language Processing (NLP) techniques have partially mitigated these challenges. Particularly, open-source Fine-tunable Pre-trained Language Models (F-PLMs), such as BERT, have shown state-of-the-art performance in various NLP tasks, including text classification and question answering, due to their ability to learn contextual meanings bidirectionally [6,7]. Similarly, localized Korean models like KoELECTRA have also been developed and applied successfully for disaster message analysis [8]. However, applying F-PLMs to disaster news classification requires extensive labeled data. Manually labeling large volumes of news data is costly and time-consuming [9]. Therefore, efficient labeling strategies combined with context-aware language models are necessary to handle large-scale, multi-type disaster news classification effectively.

To overcome these limitations, recent studies have focused on the use of Large Language Models (LLMs). Commercial API-served Large Language Models (C-LLMs), such as GPT-3 and GPT-4, can classify texts instantly through zero-shot classification without additional training, leveraging their extensive pre-trained knowledge bases [10,11]. For instance, recent GPT-based models accurately identified natural disaster-related articles from millions of news items and effectively extracted geographical information [12]. Although some studies have shown that automated classification performance using C-LLMs approaches human-level accuracy [13], persistent use of API-based models still poses significant cost and data sovereignty issues [14,15]. Thus, efficient strategies for utilizing C-LLMs are essential to address real-time processing demands and cost constraints.

Recently, hybrid approaches combining large-scale LLMs and smaller-scale fine-tunable language models have been proposed as a solution to these challenges [11]. The hybrid approach typically employs C-LLMs first as labeling tools to automatically annotate large-scale news corpora. Subsequently, these labeled datasets are utilized to fine-tune smaller, open-source F-PLMs. Farr et al. [11] demonstrated that this hybrid method achieved higher accuracy than pure LLM-based classification, even when expert validation was limited to a small subset of the data. Additionally,

ensemble methods combining multiple F-PLMs have been reported to yield more stable and generalizable performance compared to single-model approaches [16].

In this study, we propose a hybrid news classification framework for Korean regional disaster news that integrates automated labeling using C-LLMs with an ensemble classifier of transformer-based F-PLMs. Specifically, we first use state-of-the-art C-LLMs to automatically label Korean regional news articles with disaster categories. Next, we fine-tune two selected F-PLMs using the labeled data and combine them into an ensemble classification model. The proposed hybrid model is designed to accurately and rapidly identify various disaster types, including natural disaster and man-made disasters, extracting meaningful information from large-scale regional news datasets. We further evaluate the effectiveness of our proposed approach using actual Korean disaster data, comparing the model's performance against real-world disaster events.

2. Data

2.1. News Dataset

This study utilized news headline data collected from representative regional newspapers in 16 metropolitan cities and provinces across South Korea. The analyzed regions include Seoul, Incheon, Daejeon, Daegu, Busan, Gwangju, Ulsan, Gyeonggi-do, Gangwon-do, Chungcheongbuk-do, Chungcheongnam-do, Jeollabuk-do, Jeollanam-do, Gyeongsangbuk-do, Gyeongsangnam-do, and Jeju-do. The selected regional newspapers are the Seoul Shinmun, Incheon Ilbo, Daejeon Ilbo, Maeil Shinmun, Busan Ilbo, Gwangju Ilbo, Ulsan Maeil, Gyeonggi Ilbo, Gangwon Ilbo, Chungbuk Ilbo, Chungcheong Today, Jeonbuk Domin Ilbo, Jeonnam Ilbo, Gyeongbuk Ilbo, Gyeongnam Domin Ilbo, and Halla Ilbo. The complete dataset was directly acquired from the Korea Press Foundation, covering the period from January 1, 2011, to December 31, 2024, totaling 2,771,589 news articles.

For disaster type classification, we adapted the disaster categories defined by Korea's Framework Act on the Management of Disasters and Safety. Specifically, we grouped disasters into two main categories: natural disasters and man-made disasters. The natural disaster category includes 16 types: typhoon, flood, strong wind, wind and waves, tidal wave, landslide, drought, earthquake, sandy dust, green algae bloom, red tide, heavy snowfall, cold wave, heat wave, lightning, and volcanic activity. The man-made disasters category consists of 15 types: fire, explosion, wildfire, collapse, road accidents, train accidents, aviation accidents, marine accidents, radioactive accidents, chemical accidents, fine dust, environmental pollution incidents, the paralyzation of the national core infrastructure, infectious diseases, and contagious animal diseases.

To establish objective criteria for actual disaster occurrences, we collected official dates of activation for two national disaster response bodies: the Central Disaster and Safety Countermeasures Headquarters (CDSCH) and the Central Disaster Management Headquarters (CDMH). CDSCH is the primary government body responsible for coordinating and overseeing national disaster prevention, preparedness, response, and recovery activities. In contrast, CDMH is established by responsible agencies for on-site disaster management and recovery operations.

For man-made disasters, we obtained the activation dates directly from the official disaster yearbooks. For natural disasters, dates were collected by reviewing official press releases from the Ministry of the Interior and Safety. As a result, a total of 92 activation dates were identified from 2011 to 2024, comprising 43 activations of CDSCH and 49 activations of CDMH.

2.2. Models Used

This study employed two types of language models to accurately classify disaster types from news headlines (Table 1). The first type includes F-PLMs. Specifically, we utilized BERT and KoELECTRA as representative F-PLMs. BERT is a Transformer-based model that captures contextual meanings by simultaneously considering the entire sentence context in a bidirectional manner [6]. BERT was pre-trained on diverse multilingual datasets, such as Wikipedia, enabling robust and effective handling of various languages, including Korean. Particularly, BERT accurately identifies critical keywords within short texts, such as news headlines, thus ensuring reliable classification performance.

KoELECTRA is a Korean-adapted version of ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately). Instead of BERT's Masked Language Modeling (MLM), KoELECTRA uses the Replaced Token Detection (RTD) method, significantly enhancing training efficiency [17]. KoELECTRA demonstrated outstanding performance on Korean text classification benchmarks [18]. All F-PLM training and inference tasks in this study were performed using a single NVIDIA A40 GPU.

The second type comprises C-LLMs. We selected GPT-3.5-Turbo (OpenAI) and Gemini-2.0 Flash-Lite (Google) as representative C-LLMs. C-LLMs are pre-trained on vast-scale data and capable of understanding and classifying input texts without additional training. Both GPT-3.5-Turbo and Gemini-2.0 Flash-Lite models chosen for this research are well-suited for short-text inputs, making them particularly suitable for classifying disaster-related news headlines. Moreover, the API costs for these two models are approximately 70–95% lower compared to higher-end language models, making them highly cost-effective for large-scale data analysis [19,20].

In our hybrid approach, we first applied C-LLMs to automatically label the news data. Subsequently, we fine-tuned two F-PLMs (BERT and KoELECTRA) using these automatically labeled data. Finally, we combined these two fine-tuned models into a single ensemble classifier, achieving both high classification accuracy and cost efficiency.

2.3. Construction of Training and Evaluation Dataset

In this study, we initially applied a keyword-based rule classification method [21] to approximately identify disaster-related news headlines within our dataset of 2,771,589 headlines. The keyword-based rule classification involves identifying the frequency of predefined keywords (approximately 1,500 keywords) in news headlines to categorize articles by disaster type. Fig. 1 presents the proportion of disaster-related news for each media outlet obtained through this method. The average proportion of disaster-related articles was about 3.3%, ranging from a minimum of 2.0% (Gyeongbuk Ilbo) to a maximum of 4.9% (Gwangju Ilbo). Most media outlets showed disaster-related news proportions between approximately 2.5% and 4%. This distribution indicates a highly sparse environment regarding disaster news within the total dataset. Moreover, the frequency of news varied considerably among disaster types (Fig. 2). Specifically, rare disaster types such as volcanic activity and wind and waves had fewer than a few hundred news articles. Such class imbalance can lead classification models to achieve deceptively high accuracy simply by predicting the majority class ('Non-disaster').

Additionally, since the collected news data span 14 years (from 2011 to 2024), the language and terms used in news reporting may evolve over time (concept drift). Training a model using data from only one specific time period may thus yield unstable or inconsistent classification performance.

To address these issues, we performed stratified sampling. We divided the total dataset into three periods: 2011–2015, 2016–2020, and 2021–2024. Within each period, the number of articles labeled as 'Non-disaster' was limited to a maximum of 5,000, while the maximum for each of the remaining 31 disaster types was set to 500 headlines. News headlines exceeding these limits were randomly selected. Through this stratified sampling approach, we ultimately constructed a balanced training dataset comprising approximately 55,000 news headlines (Fig. 3). This balanced dataset effectively mitigates class imbalance and addresses potential temporal changes in news reporting styles.

3. Proposed Hybrid Classification Framework

3.1. Data Labeling Using Commercial API-served Large Language Models (C-LLMs)

In this study, we performed disaster-type labeling on 56,563 news headlines selected through stratified sampling. Two C-LLMs—GPT-3.5-Turbo (OpenAI) and Gemini-2.0 Flash-Lite (Google)—were utilized for this labeling process. We first prepared a prompt instructing the models to classify each news headline into the most appropriate category among 32 disaster types (including 'Non-disaster'). Then, we provided the prompt to GPT-3.5-Turbo and Gemini-2.0 Flash-Lite separately and obtained classification results from each model.

Upon evaluating the agreement between the two models, the classification results matched for 45,765 out of the 56,563 headlines (80.9%). For these cases, the automatically labeled results were directly accepted as the final labels (Fig. 4). The remaining 10,798 headlines (19.1%), for which the two models' classifications differed, were manually reviewed and labeled by the research team to finalize the labels. This hybrid labeling method significantly reduced manual labeling effort while maintaining data accuracy. Consequently, our approach demonstrates an effective strategy for efficient, high-quality dataset construction by substantially minimizing human annotation effort.

3.2. Fine-tuning of Open-source Fine-tunable Pre-trained Language Models (F-PLMs)

In this study, we optimized disaster news classification performance by fine-tuning two F-PLMs: BERT and KoELECTRA. To achieve optimal performance, we systematically explored two critical hyperparameters: learning rate (LR) and training epochs. All experiments were conducted using stratified 5-fold cross-validation. The evaluation metric was the Macro F1 score, defined as the unweighted mean of class-specific F1 scores, which effectively accounts for class imbalance by equally weighting each disaster category regardless of sample size.

Appropriate selection of learning rate is crucial when fine-tuning F-PLMs. An excessively high learning rate can rapidly alter pre-trained representations, leading to model divergence, while an excessively low learning rate can slow down the training process, resulting in underfitting [6, 22]. Based on recommended learning rate ranges (1e−5 to 5e−5) from prior literature, we tested a total of seven candidate values. Both BERT and KoELECTRA achieved the highest Macro-F1 scores at a learning rate of 3e−5 (BERT: 0.9439, KoELECTRA: 0.9493). In contrast, performance declined notably at extremely low (5e−7) or high (1e−4) learning rates (Fig. 5, (a), (b)).

We further evaluated model performance by varying the training epochs from 1 to 5. BERT initially achieved a Macro-F1 score of 0.9247 at Epoch 1 and reached its highest performance of 0.9448 at Epoch 5. KoELECTRA started with a Macro-F1 score of 0.9371 at Epoch 1, showing a gradual increase and converging at 0.9519 by Epoch 5 (Fig. 5, (c), (d)). However, the incremental improvement at Epoch 5 for both models was marginal (less than 0.1%), despite approximately 25%

more computational resources being required. Thus, considering performance and computational efficiency, we selected Epoch 4 for BERT and KoELECTRA as the optimal training epochs.

Additionally, we combined the individually optimized BERT and KoELECTRA models to further improve overall performance. To achieve this, we employed a soft-voting ensemble method by averaging the class probability vectors (32 classes) generated by each model. BERT (MLM, WordPiece tokenizer) and KoELECTRA (RTD, SentencePiece tokenizer) differ in pre-training objectives and tokenization methods, resulting in decorrelated error patterns. Combining such models with decorrelated prediction behaviors reduces prediction variance and enhances recall for rare classes [23, 24]. Consequently, our proposed soft-voting ensemble model provides more stable and balanced classification performance compared to single-model approaches.

3.3. Performance Comparison of C-LLMs and F-PLMs

This study compared the classification performance of six disaster classification models using the constructed news dataset (Table 2). The evaluated models included one keyword-based rule classification model, two F-PLMs (BERT and KoELECTRA), two C-LLMs (GPT-3.5-Turbo and Gemini-2.0 Flash-Lite), and one soft-voting ensemble combining the two F-PLMs. Model performance was assessed using the Macro-F1 score, employing a stratified 5-fold cross-validation approach.

Among the evaluated models, Gemini-2.0 Flash-Lite (C-LLM) achieved the highest Macro-F1 score of 0.9817, whereas GPT-3.5-Turbo showed a relatively lower performance of 0.8178. The F-PLMs—KoELECTRA and BERT—achieved high Macro-F1 scores of 0.9519 and 0.9441, respectively, significantly outperforming both the keyword-based rule classification (0.8920) and GPT-3.5-Turbo. Moreover, the soft-voting ensemble of the two F-PLMs further improved classification performance, achieving a Macro-F1 score of 0.9589, surpassing the results of the individual F-PLMs.

An important implication of these results is that a hybrid approach—using C-LLMs primarily for data labeling and cost-effective F-PLMs for the actual classification service—is highly efficient. Specifically, the hybrid labeling strategy (§3.1) employed in this study effectively minimized the use of C-LLMs, thereby obtaining high-quality labeled data. The fine-tuned F-PLMs and their ensemble, trained on these labeled data, demonstrated performance comparable to that of the best-performing C-LLM, ensuring both high accuracy and cost-efficiency in real-world operational contexts.

4. Validation of the Proposed Model with Real-World Disaster Events

To validate the practical applicability of our proposed hybrid news classification framework, we classified a total of 2,771,589 news headlines collected between 2011 and 2024 by disaster type. We visualized these classification results in a time-series format and compared them against actual disaster events, represented by activation dates of Korea's CDSCH and CDMH, to visually assess the model's performance (Fig. 6).

The peaks in classified news article frequency closely corresponded with actual large-scale disaster occurrences. This alignment was especially evident for major disaster types, including typhoons, floods, earthquakes, heat waves, fires, forest fires, ship accidents, and infectious disease outbreaks. For instance, news coverage spiked dramatically during the 2015 MERS outbreak and the 2020 COVID-19 pandemic, accurately aligning with activation periods of the CDSCH. Similarly, typhoon and flood events demonstrated clear seasonal patterns, with significant news coverage peaks precisely

matching actual disaster incidents, such as Typhoon Khanun in 2023, Typhoon Hinnamnor in 2022, and Typhoon Soulik in 2018.

These findings confirm that the proposed hybrid classification framework effectively captures real-world disaster events in real time, providing accurate and timely information valuable for disaster management and response operations.

5. Conclusion and Discussion

This study proposed and validated an efficient hybrid classification framework for rapidly detecting and classifying disaster-related news from Korean regional news datasets. The proposed framework leverages C-LLMs for automated labeling, efficiently constructing high-quality labeled datasets. Subsequently, two F-PLMs, specifically BERT and KoELECTRA, were fine-tuned using this labeled dataset. We further enhanced classification performance by combining these two F-PLMs into a soft-voting ensemble classifier.

Specifically, two representative C-LLMs—GPT-3.5-Turbo (OpenAI) and Gemini-2.0 Flash-Lite (Google)—were employed for the automated labeling of 56,563 news headlines. The labeling results from these models showed an agreement rate of approximately 80.9%, significantly reducing manual labeling effort by a factor of five while maintaining high accuracy. The fine-tuned F-PLMs models, BERT and KoELECTRA, achieved Macro-F1 scores of 0.9441 and 0.9519, respectively. Moreover, the soft-voting ensemble of these two models further improved the classification performance to a Macro-F1 score of 0.9589. These results highlight the effectiveness of our cost-efficient hybrid approach, achieving performance comparable to the high-performance yet costly C-LLMs.

To validate the practical applicability of the proposed hybrid classification framework, we classified the complete news dataset comprising 2,771,589 news headlines collected from 2011 to 2024. The resulting disaster-type classifications were visualized in a time-series format and compared against the official activation dates of national disaster response organizations (CDSCH and CDMH). Our analysis demonstrated that news coverage peaks closely aligned with the occurrence dates of major disasters, such as typhoon, flood, earthquake, heat wave, fire, wildfire, marine accidents, and infectious disease outbreaks. These findings demonstrate that the proposed model, utilizing news data as a single source, can serve as a valuable supplementary information source across a broad spectrum of disaster types, effectively supporting situational awareness and informed decision-making in disaster management and response operations.

In practice, the hybrid classification framework presented in this study is actively employed in disaster management settings in Korea. Firstly, the National Disaster Management Research Institute (NDMI) utilizes this model for real-time monitoring of disaster-related news through a dedicated web page (Fig. 7) [25]. Additionally, this system is also utilized by the Disaster and Safety Management Centers of the Ministry of the Interior and Safety. Secondly, the time-series data on disaster-related news trends generated by our model serve as one of the essential references for the Ministry of the Interior and Safety in establishing annual investment priorities for national disaster and safety management programs.

Nevertheless, this study has certain limitations and areas requiring further research. Firstly, the analysis utilized only Korean-language regional news data from South Korea. Consequently, additional validation and generalization tests are needed to apply this framework internationally or in other linguistic contexts. Secondly, the current classification model focuses only on short texts, such as news headlines. Future studies should develop extended models incorporating full news articles to capture deeper contextual information and semantic nuances.

Despite these limitations, our proposed hybrid approach provides a practical and cost-effective methodology suitable for disaster management organizations and university research labs with limited budgets, facilitating broader adoption of artificial intelligence technologies. By effectively combining cost-efficient F-PLMs and minimally relying on expensive C-LLMs, this study significantly contributes toward improving the practical application and quality of AI-driven disaster management and response.

6. Funding

This work was supported by the National Disaster Management Research Institute (NDMI), project number NDMI-PR-2025-04-01.

7. Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

References

[1] A. Chauhan, A.L. Hughes, Providing online crisis information: an analysis of official sources during the 2014 Carlton Complex wildfire, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, 2017, pp. 3151–3162. https://doi.org/10.1145/3025453.3025627.

[2] N. Kapoor, Role of media in disaster management, J. Adv. Res. Soc. Sci. Humanit. 1 (2015) 5–8.

[3] T. Pisal, V. Jadhav, The use of big data to improve disaster response and preparedness efforts, Int. J. Adv. Res. Sci. Commun. Technol. 5 (2024) 1–8. https://doi.org/10.48175/IJARSCT-15086.

[4] M.M. De Brito, C. Kuhlicke, A. Marx, Near-real-time drought impact assessment: a text-mining approach on the 2018/19 drought in Germany, Environ. Res. Lett. 15 (2020) 1040a9. https://doi.org/10.1088/1748-9326/aba4ca.

[5] H. Bailon, K. Boersma, C. Orellana-Rodriguez, M. Van den Homberg, Framing of disaster impact in online news media: a case study from Malawi on flood risk management, Front. Commun. 10 (2025) 1519357. https://doi.org/10.3389/fcomm.2025.1519357.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, pp. 4171–4186.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, V. Stoyanov, RoBERTa: a robustly optimized BERT pre-training approach, arXiv preprint arXiv:1907.11692 (2019).

[8] Y. Lee, S. Song, Do language models flexibly process language expressions? An evaluation using Korean disaster messages, Lang. Facts Perspect. 62 (2024) 229–255.

[9] F. Alam, U. Qazi, M. Imran, F. Ofli, HumAID: human-annotated disaster incidents data from Twitter with deep-learning benchmarks, arXiv preprint arXiv:2104.03090 (2021).

[10] OpenAI, GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[11] D. Farr, N. Manzonelli, I. Cruickshank, J. West, RED-CT: a systems design methodology for using LLM-labeled data to train and deploy edge classifiers for computational social science, arXiv preprint arXiv:2408.08217 (2024).

[12] F. Sufi, M. Alsulami, AI-driven global disaster intelligence from news media, Mathematics 13 (2025) 1083. https://doi.org/10.3390/math13071083.

[13] L. Bojić, O. Zagovora, A. Zelenkauskaite, V. Vuković, M. Čabarkapa, S. Veseljević Jerković, A. Jovančević, Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm, Sci. Rep. 15 (2025) 11477. https://doi.org/10.1038/s41598-025-96508-3.

[14] I.J. Cruickshank, L.H.X. Ng, Prompting and fine-tuning open-sourced large language models for stance classification, arXiv preprint arXiv:2309.13734 (2024).

[15] R. Zhang, Y. Li, Y. Ma, M. Zhou, L. Zou, LLMaAA: making large language models as active annotators, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 13088–13103. https://doi.org/10.18653/v1/2023.findings-emnlp.872.

[16] H. Zhang, M.O. Shafiq, Survey of transformers and towards ensemble learning using transformers for natural language processing, J. Big Data 11 (2024) 25.

[17] Y. Heo, J. Lee, J. Kim, K. Park, KoELECTRA: pre-training a large-scale Korean language representation generator with replaced token detection, arXiv preprint arXiv:2104.06940 (2021).

[18] S.-T. Park, Y. Kim, J. Oh, et al., KLUE: Korean language understanding evaluation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.

[19] OpenAI, GPT-3.5-Turbo API pricing, 2024. https://openai.com/pricing (accessed 14 June 2025).

[20] Google, Gemini 2.0 Flash-Lite API pricing, 2024. https://ai.google.dev (accessed 14 June 2025).

[21] E.H. Shin, D.W. Kim, J.H. Chung, S.R. Chang, Development of a method for measuring social interest index on disaster using news data, J. Korean Soc. Saf. 38 (5) (2023) 27–35.

[22] S. Jastrzębski, Z. Kenton, M. Kim, D. Arpit, N. Ballas, A. Fischer, et al., The break-even point on optimization and generalization in deep learning, arXiv preprint arXiv:2002.09582 (2020).

[23] T.G. Dietterich, Ensemble methods in machine learning, in: B.S. Jones, R.Z. Smith (Eds.), Multiple Classifier Systems, LNCS, vol. 1857, Springer, 2000, pp. 1–15.

[24] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CRC, 2012.

[25] National Disaster Management Research Institute, Real-time Monitoring of Disaster-related News. https://rscanner.ndmi.go.kr/, 2025 (accessed 30 June 2025).

Title page

Manuscript Title
LLM-Assisted Labeling and Transformer Model Ensemble for Disaster-Related News Classification in Korean Regional Media (2011–2024)

Authors
Eunhye Shin
National Disaster Management Research Institute, 365, Jongga-ro, Jung-gu, Ulsan, South Korea
Email: shinehy@korea.kr
ORCID: https://orcid.org/0009-0009-0649-3049

Gahee Lee
National Disaster Management Research Institute, 365, Jongga-ro, Jung-gu, Ulsan, South Korea
Email: gh0302@korea.kr

Do-Woo Kim (Corresponding Author)
National Disaster Management Research Institute, 365, Jongga-ro, Jung-gu, Ulsan, South Korea
Email: dow1112@korea.kr

Acknowledgment
This research was supported by the National Disaster Management Research Institute (NDMI), Korea (Project No. NDMI-PR-2025-04-01).

Conflict of Interest
The authors declare that they have no conflict of interest.

Keywords
large language model, pre-trained language model, disaster-related news classification, disaster management