Development of AI-Driven Korean-Language Rumor Detection Model for Radiological Emergencies

Jeon Inyoung
Korea Institute of Nuclear Safety
k137jiy@kins.re.kr

Keywords: rumor detection, AI, machine learning, radiological emergency

1. Introduction

In the digital age, misinformation and rumors spread rapidly through online platforms, negatively influencing public perception and decision-making. Detecting and mitigating false information is crucial, especially for sensitive topics such as nuclear safety and environmental issues [1]. This study focuses on developing an AI-driven rumor detection system for analyzing Korean-language text related to Fukushima-related discussions.

Leveraging advances in Natural Language Processing (NLP) and deep learning, a BERT-based classification model was employed to distinguish between rumors and non-rumors in Korean text. The model is fine-tuned on a labeled dataset, enabling it to recognize linguistic patterns associated with misinformation.

The proposed system follows a structured pipeline, including data preprocessing, text tokenization, model training, and evaluation. To address potential class imbalances in the dataset, up-sampling technique was added to enhance model robustness. The training process optimized performance using supervised learning techniques and early stopping mechanisms to prevent overfitting.

2. Methods and Results

In this section some of the techniques used to develop rumor detection model are described. The proposed system follows a structured pipeline, including data preprocessing, text tokenization, model training, and evaluation.

2.1 Data Preprocessing

The dataset used for this study consists of Korean-language text related to Fukushima rumors, stored in a CSV file (korean_fukushima_rumors.csv). The dataset contains two primary columns (text and label): textual content to be analyzed and a binary classification indicating whether the text is a rumor (1) or not a rumor (0). The rumors included exaggerated, false, or misleading claims regarding the spread of radiation, health risks, and contamination of food or water from Japan's Fukushima nuclear disaster. Examples of rumors include claims that radiation from Fukushima has reached Korea, that the consumption of certain domestic foods can lead to radiation exposure, or that the effects of radiation released from Fukushima nuclear plants to South Korea can cause extreme health issues like cancer, mutations, or death.

On the other hand, the non-rumors are scientifically grounded statements or clarifications that refute these exaggerated claims. These statements often explain the scientific facts about radiation dispersion, safety standards, and the lack of significant risks from Fukushima-related radiation.

Since real-world datasets often suffer from class imbalances, up-sampling technique was employed to increase the representation of the minority class. The dataset was split into 80% training data and 20% validation data.

2.2 Text Tokenization

To process text data for deep learning, KoBERT, a Korean-language adaptation of BERT was utilized [2]. Tokenization was performed using the AutoTokenizer from the Hugging Face transformers library. The tokenized text was formatted into tensor-based inputs for compatibility with the deep learning model.

2.3 Model Training

The training process was conducted using the Hugging Face Trainer API, which provides an efficient way to fine-tune transformer-based models. The model selected for this task was KoBERT (monologg/kobert), a variant of BERT pretrained specifically for Korean text. The training procedure involved supervised learning using labeled data, where the model learned to classify text as either rumor (1) or not a rumor (0).

The model's classification head consists of a fully connected layer applied to the [CLS] token representation, followed by a softmax activation function to generate class probabilities. The model was optimized using cross-entropy loss, which is well-suited for binary classification tasks.

After training, the best-performing model and tokenizer were saved for later use. The saved model can be reloaded and used for inference on new text data. This ensures that the system remains scalable and deployable for real-world misinformation detection applications.

2.4 Model Evaluation

The author compiled the Fukushima nuclear rumor dataset by collecting real-world rumors and factual statements from social media, news sources, and discussion forums following the Fukushima nuclear

accident. Each text sample was manually labeled as either "rumor" (1) or "not a rumor" (0) based on its verifiability from credible sources such as scientific reports from ICRP, IAEA, or NCRP. The dataset used for model comprised 100 statements, with 54% classified as accurate and 46% as misinformation. The KoBERT model was fine-tuned in the Google Colab environment, utilizing available GPU acceleration. After training the KoBERT-based model, its performance was evaluated on the validation set, using several key metrics to assess its ability to classify text as rumor or not a rumor. The model demonstrated strong performance across all evaluation criteria, particularly in distinguishing between rumors and non-rumors in Korean-language text. The results are summarized in Table 1.

Table 1: Technical Performance of the Model

Metric	Value
Accuracy	0.810
Precision	0.750
Recall	0.900
F1 Score	0.818

Study results suggest that the model is well-suited for the task of rumor detection in Korean-language text, demonstrating a high level of accuracy, precision, recall, and F1-score.

In addition, the model's inference capabilities were evaluated by testing it on a set of previously unseen Korean text samples. The rumor detection function accurately classified new instances, providing both the classification result (Rumor or Not a Rumor) and a confidence score (See Table 2)

Table 2: Model's Inference Capability (Examples)

Input Text	Output
"후쿠시마 원전 사고로 서울의 수도물도 방사능에 오염되었다"	Rumor (0.831)
"한국 연안 해산물은 방사능에 오염되지 않았다"	Not a Rumor (0.712)

These results confirm that the model is capable of delivering reliable predictions in real-world settings.

3. Conclusions

Rumors can trigger unnecessary anxiety, confusion, and fear among the public, which may lead to harmful actions. Furthermore, emergency services and responders may experience an increased workload as they need to handle false reports and misinformation, diverting their resources away from emergency situations. In this regard, the development and implementation of a rumor detection system is crucial to detect and analyze rumors efficiently and effectively [4].

The model's strong performance in terms of accuracy, precision, recall, and F1-score indicated that the fine-tuned KoBERT model is highly effective for classifying Korean-language text related to Fukushima rumors. Historical radiological emergencies such as Fukushima nuclear disaster (2011), the Chernobyl disaster (1986), and the Three Mile Island accident (1979) have shown the significance of immediate and accurate communication in safeguarding public health and safety. By leveraging transfer learning, we are able to take advantage of KoBERT's pretrained language knowledge and adapt it to the specific task of rumor detection in the case of potential radiological emergencies. Future research should focus on developing real-time rumor detection systems that can quickly identify and analyze potential misinformation on social media and online platforms.

REFERENCES

[1] Paek, H. J., & Hove, T. (2019). Effective strategies for responding to rumors about risks: The case of radiation-contaminated food in South Korea. Public relations review, 45(3), 101762.

[2] Anggrainingsih, R., Hassan, G.M., & Datta, A. (2022). Evaluating BERT-based Pre-training Language Models for Detecting Misinformation. ArXiv, abs/2203.07731.

[3] Choi, D. J., Oh, H. C., Chun, S. L., Kwon, T. Y., & Han, J. Y. (2022). Preventing rumor spread with deep learning, Expert Systems with Applications, Volume 197

[4] Agarwal P., Aziz R.A., Zhuang J., Interplay of rumor propagation and clarification on social media during crisis events - A game-theoretic approach, European Journal of Operational Research 298 (2) (2022) 714–733.