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

In Young Jeon, PhD
k137jiy@kins.re.kr
Korea Institute of Nuclear Safety

Background

- Rumors can trigger unnecessary anxiety, confusion, and fear among the public, which may lead to harmful actions s shown in historical radiological disasters (e.g., Fukushima 2011, Chernobyl 1986, Three Mile Island 1979)
- 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 or evaluation system is crucial to detect and analyze rumors efficiently and effectively

Purpose

- This study developed AI-driven Korean-language rumor detection model for radiological emergencies, particularly for Fukushima nuclear accident

Methods and Results

- Preparation of a structured dataset
  - Korean-language rumors related to the Fukushima nuclear accident to test the AI-driven rumor detection system of this study (See Table1)
  - Collecting and labeling Korean-language texts from news articles, social media, and discussion forums.
  - The rumors including exaggerated, false, or misleading claims regarding the spread of radiation, health risks, and contamination of food or water from Japan's Fukushima nuclear accident.
  - Non-rumors scientifically grounded statements or clarifications that refute exaggerated claims based on their verifiability from credible sources such as scientific reports from ICRP, IAEA, or NCRP.

Methods and Results

- The dataset initially contained 50 statements, later expanded with 50 AI-generated statements to enhance robustness.
- Of the total data, 54% were classified as accurate (label 0), while 46% were identified as misinformation (label 1).

Table 1. Example of dataset of Korean-language rumors for Fukushima nuclear accident

	A	B
1	text	label
2	"후쿠시마 방사능이 한국 바다에 도달했다!"	1
3	"일본은 후쿠시마 오염수를 안전하게 방류하고 있다."	0
4	"후쿠시마 방사능이 바람을 타고 한국 전역을 오염시키고 있다."	1
5	"과학적 연구에 따르면 후쿠시마 방사능 수치는 안전한 수준이다."	0
6	"편서풍 때문에 우리나라에 방사능 피해가 없다"	0
7	"중국에서 한국을 거쳐 일본으로 가는 북서·북풍계열 바람이 불고 있다"	0
8	"기류의 방향 등을 볼때 일본 동북지방에서 유출 가능성이 있는 방사능이 우리나라 쪽으로 이동할 가	0
9	"일본 열도에서 유출된 방사능이 한반도에 상륙한다"	1
10	"바람 방향이 한국쪽으로 바뀌면서 방사능이 한국에 상륙할?것이다"	1
11	"비가 내릴 경우 처음 24시간 동안 문과 창문을 모두 닫고 집안에 머물러 있어야 한다"	1
12	"가랑비에라도 노출될 경우 화상을 입고, 대머리가 되거나 심지어 암에 걸릴 수도 있다"	1
13	"방사성물질이 태평양 방향으로 날아가게 되는 경우 많은 바닷물에 의해 그 농도가 희석이 된다"	0
14	"바닷물에 포함된 방사성동위원소가 정밀한 분석기술을 통하여 검출될 수도 있으나 희석되는 바닷물	0
15	"이번 후쿠시마 원전 폭발로 방출된 방사성 물질 중 가장 유해한 것은 세슘과 방사성요오드이다"	0
16	"세슘의 축적량이 절반으로 줄어드는 반감기가 30년에 달해 피폭 후유증도 오래 간다"	0
17	"방사성요오드는 역시 호흡으로 유입돼 갑상선에 축적돼 감마선이나 베타선을 방출한다"	0
18	"방사능 피폭에 녹차의 타닌과 비타민, 천일염이 도움이 된다"	1
19	"해바라기를 심어놓으면 방사능 물질을 흡수한다"	1
20	"방사능 피폭에 혈액순환에 레드와인, 소주, 보드카가 도움이 된다"	1

- Text Tokenization
  - To process text data for deep learning, KoBERT, a Korean-language adaptation of BERT was utilized.
  - Tokenization was performed using the AutoTokenizer from the Hugging Face transformers library.
- Model Training
  - The training process was conducted using the Hugging Face Trainer API, which provides an efficient way to fine-tune transformer-based models, updating their parameters.
  - 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).
- Model Evaluation
  - The dataset split into training (80%) and validation (20%) sets
  - After training the KoBERT-based model, its performance was evaluated on the validation set
  - The model demonstrated strong performance across all evaluation criteria (See Table 2).

Methods and Results

Table 2. Technical Performance of the Study Model

Metric	Value
Accuracy	0.810
Precision	0.750
Recall	0.900
F1 Score	0.818

- In addition, the model's inference capabilities were evaluated by testing it on a set of previously unseen Korean rumor samples for Fukushima nuclear accident.
- The evaluation results demonstrated the model's ability to distinguish misinformation from facts. (see Table 3)

Table 3. Inference Capabilities of the Study Model

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

Discussion & Conclusion

- This study presented the first AI-driven Korean language rumor detection model for radiological emergencies, particularly for Fukushima nuclear accident
- 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.
- Scientific uncertainties were sometimes misclassified as rumors, highlighting the need for further model refinement
- 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

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