Analyzing the SMR Discourse through Hate Speech Classification and Topic Modeling : A Multi-layered Approach to Unmasking Online Conflict

Jisong Jeong [a], Philseo Kim [a,*], and Young Bae [b,*]
[a] Division of Advanced Nuclear Engineering, Pohang University of Science and Technology (POSTECH), 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, Republic of Korea
[b] The Division of Humanities and Social Science, Pohang University of Science and Technology (POSTECH), 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, Republic of Korea
[*] *Co-Corresponding author: philseokim@postech.ac.kr, youngbae@postech.ac.kr*

***Keywords :** Small Modular Reactor, Hate Speech Detection, Multi-label Text Classification ,Topic Modeling*

1. Introduction

Small Modular Reactors (SMRs), emerging as a next-generation energy source, face the dual challenges of technological opportunity and social acceptance. The online public sphere serves as a key space for shaping public perception and opinion on SMRs, while simultaneously acting as an arena where social conflicts are amplified. Grounded in the concern that the scientific issue of SMRs can devolve from a subject of rational debate into an outlet for pre-existing social and political conflicts, this study aims to analyze the SMR discourse through online comment data. This study aims to reveal the hidden dynamics and structure of conflict within the SMR online discourse by constructing a deep learning-based hate speech classification model and applying topic modeling.

2. Methods and Results

2.1 Data and Preprocessing

The analysis was conducted on a dataset of approximately 240,583 online comments from Naver news articles related to SMRs. To enhance analytical accuracy, custom dictionaries for synonyms and stopwords relevant to the research topic were created. The Komoran morphological analyzer was used for text processing, with a user dictionary applied to ensure that key terms such as ‘SMR’ and ‘Small Modular Reactor’ were treated as single tokens, thereby improving the quality of the text data.

2.2 Hate Speech Classification Model: Construction and Optimization

To classify nine types of hate speech (‘Origin’, ‘Physical’, ‘Politics’, ‘Profanity’, ‘Age’, ‘Gender’, ‘Religion’, ‘Race’, ‘Not Hate Speech’), an augmented version of the ‘jeanlee/kmhas_korean_hate_speech’ dataset, a benchmark for Korean hate speech detection available on Hugging Face, was used as the training data. For the model architecture, a multi-label classification model was built based on the ‘klue/roberta-large’ pretrained language model, which is known for its high performance in Korean natural language processing. Through hyperparameter tuning, the model achieved its best performance with an F1 Score of 0.7764 under the conditions of 5 epochs and a learning rate of 1e-5.

Furthermore, to maximize the model's potential, an optimal classification threshold was determined using the validation set. The analysis revealed that the highest F1 Score of 0.7827 was achieved at a threshold of 0.35, which is lower than the default value of 0.5. This indicates that the model more effectively identifies hate speech when configured with higher sensitivity. Consequently, this optimal threshold was applied for the final comment classification.

Table I: models on the benchmark set.

	Model	Epochs	Learning Rate	Threshold	F1 Score	Accuracy
1	KoELECTRA-base	3	1e-5	0.5	0.7375	0.8162
2		5	1e-5	0.5	0.7304	0.8118
3	RoBERTa-large	3	1e-5	0.5	0.7748	0.8228
4		**5**	**1e-5**	**0.5**	**0.7764**	**0.8220**
5		**5**	**1e-5**	**0.35**	**0.7827**	**0.8153**
6	RoBERTa-large	5	2e-5	0.5	0.7676	0.8164
7		5	5e-6	0.5	0.7692	0.8200

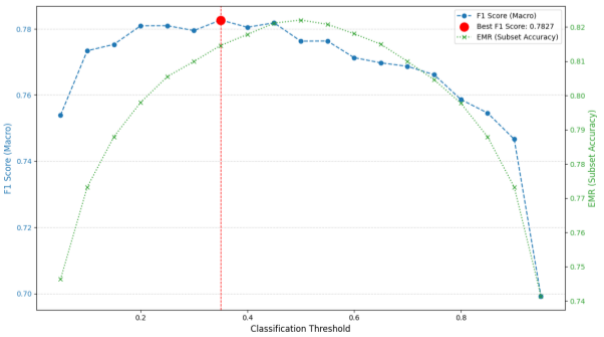

Figure 1. F1 score & EMR classification Threshold

2.3 Quantitative Overview of the SMR Discourse

Applying the optimized model to the entire dataset, 70.6% of the comments were classified as ‘Non-Hate

Speech', while 29.4% were classified as 'Hate Speech'. An analysis of the hate speech types revealed that 'Politics' was the overwhelmingly dominant category, accounting for 74.6% of all identified hate speech instances. This was followed by 'Profanity' (11.8%) and 'Origin' (regional discrimination, 9.0%), suggesting that conflict within the SMR discourse is more deeply intertwined with politics than any other factor.

Table II: Ratio of Hate Speech to Non-Hate Speech

Comment Type	Count (Ratio)
Non-Hate Speech	169,886 (70.6%)
Hate Speech	70,697 (29.4%)

Table III: Distribution of Hate Speech Type

Hate Speech Type	Count (Ratio)
Politics	56,658 (74.6%)
Profanity	8,994 (11.8%)
Origin	6,836 (9%)
Age	1,385 (1.8%)
Physical	1,186 (1.6%)
Gender	683 (0.9%)
Religion	132 (0.2%)
Race	44 (0.1%)

2.4 Discourse Analysis: Topic Modeling

2.4.1 The Multifaceted Topics of Non-Hate Speech

The topic modeling of 'Non-Hate Speech' comments revealed that the public discourse on SMRs is structured around three distinct themes: 'Critique from a Citizen/Shareholder Perspective', 'Political and Media Discourse', and 'Debate on Technology and Policy'. This demonstrates that the SMR discourse is not a simple binary of pro versus con, but rather a complex space where various actors with different viewpoints and interests participate.

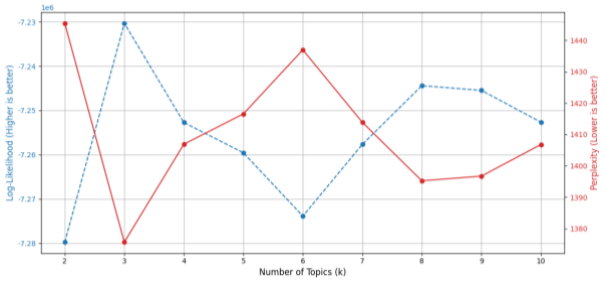

Fig. 2. Metrics for Optimal Number of Topics (Non-Hate Speech)

Table IV: Results of topic modeling

Topic	Keywords	# of comments	Contents
1	people, person, budget, Democratic Party, human, government, Doosan, country, tax, opinion, voice, cut, shareholder	54887	Critical Public Opinion on Corporate Management and Social Issues
2	president, administration, article, Yoon Suk-yeol, USA, investment, Republic of Korea, public, economy, Ahn Cheol-soo, company, Moon Jae-in, Korea	49898	Formation of Political and Media Discourse around SMRs
3	nuclear power plant, country, energy, technology, electricity, world, nuclear power, solar power, generation, policy, China, export, safety	53169	Policy Debate on SMR Technology, Safety, and Future

2.4.2 The Multi-layered Structure of 'Origin' Hate Speech

To conduct an in-depth analysis, the 'Origin' category—the type of hate speech hypothesized to be most directly related to the SMR issue—was closely examined. The results of LDA Topic Modeling revealed that this category is not a monolithic form of regional animosity but is differentiated into three distinct sub-topics: discontent with domestic politics and society at large, the geopolitical and technological competition surrounding SMRs, and ideological, regional, and xenophobic hate tied to energy policy. This finding indicates that regional conflicts related to SMRs extend beyond simple NIMBY (Not In My Backyard) phenomena and are complexly intertwined with pre-existing political and social conflict structures.

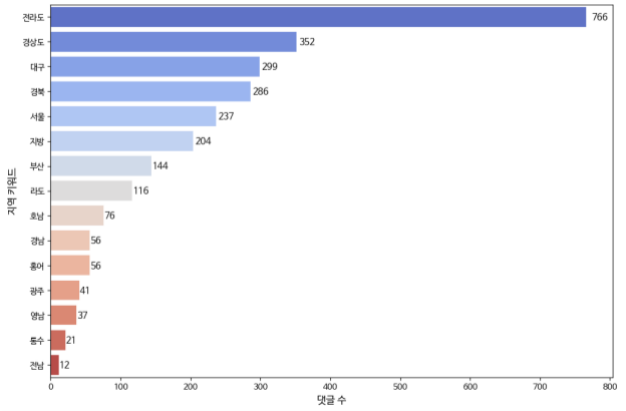

Fig. 3. Regions Most Frequently Targeted in Origin Hate Speech

2.5 Verification of Conflict Structure: N-gram Co-occurrence Analysis

To concretely verify the multi-layered structure of the 'Origin-based' hate speech discourse discovered through topic modeling, core attack phrases from each regional group were extracted using N-gram analysis.

The analysis revealed that in comments denigrating the Honam region (877 cases), a linguistic pattern of 'political stigmatization' was dominant, directly linking the region to specific political forces and energy policies with phrases such as 'President Moon Jae-in', 'Jeolla-do solar power', and 'Democratic Party Jeolla-do'.

In contrast, in comments denigrating the Yeongnam region (749 cases), the linguistic pattern primarily involved the shifting of 'geographical responsibility' related to the physical siting of nuclear facilities, featuring phrases like 'Daegu Gyeongbuk', 'nuclear plant Gyeongsang-do', and 'nuclear radioactive waste'.

This asymmetry in the logic of attack empirically demonstrates that the same SMR issue is consumed through entirely different frames depending on the socio-political context of each region. Furthermore, in comments where the Seoul Metropolitan Area-province conflict appeared (754 cases), logics of 'metropolitan egoism' and 'sacrifice of the provinces' were discovered, confirming that SMRs serve as a venue for projecting various pre-existing social conflicts.

3. Conclusions

This study identified the dual nature of the SMR discourse by analyzing approximately 240,000 online comments. First, about 70% of the total discourse consists of non-hate speech, showing a superficially healthy state where multifaceted discussions on technology, economy, and safety take place. Second, the moment the discourse turns conflictual, its content sharply skews towards criticism of specific political leanings. Third, by deeply analyzing 'Origin-based' hate speech, this study demonstrated that the SMR issue functions as a proxy war, reproducing and amplifying pre-existing political, social, and regional conflicts through the tangible problem of 'where to build'. The core of the conflict was not a lack of technical understanding, but an issue of social trust and conflict management.

These results suggest that for future SMR-related policy making and public communication strategies, it is crucial to move beyond simply promoting the technology's safety or economic benefits. The decisive challenge in securing social acceptance for SMRs lies in understanding and delicately handling the deeply rooted conflictual contexts inherent in our society.

Acknowledgements

This work was supported by Korea Hydro & Nuclear Power company through the project "Nuclear Innovation Center for Haeoleum Alliance" and the 'Human Resources Program in Energy Technology' of the Korea Institute of Energy Technology Evaluation and Planning(KETEP), which was funded by the Ministry of Trade, Industry & Energy(MOTIE. Korea). (No. RS-2024-00398425)

REFERENCES

[1] Lee, J., Lim, T., Lee, H., Jo, B., Kim, Y., Yoon, H., & Han, S. C. (2022). K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. *arXiv preprint arXiv:2208.10684*.

[2] Yoon, I. J., Han, K. D., & Kim, H. (2018). Hate Speech against Immigrants in Korea: A Text Mining Analysis of Comments on News about Foreign Migrant Workers and Korean Chinese Residents: A Text Mining Analysis of Comments on News about Foreign Migrant Workers and Korean Chinese Residents. *Asia Review*, *8*(1), 259-288.

[3] Wanniarachchi, V. U., Scogings, C., Susnjak, T., & Mathrani, A. (2023). Hate speech patterns in social media: A methodological framework and fat stigma investigation incorporating sentiment analysis, topic modelling and discourse analysis. *Australasian Journal of Information Systems*, *27*.

[4] Hlavacek, M., Cabelkova, I., Pawlak, K., & Smutka, L. (2023). Nuclear reactor at home? Public acceptance of small nuclear reactors in the neighborhood. *Frontiers in Energy Research*, *11*, 1211434.