An Engineering Approach to Datasets and Models for Language Sensitivity

Eun Huh[1], Bogyeong Kim[2], Jiseon Park[3], Chanho Park[4], Sihyeon Yang[5], Yegun Lee[6]

[1]*Computer Engineering, Hansung University, Seongbuk, Korea*
[2]*Linguistic and Cognitive Science, Hankuk University of Foreign Studies, Yongin, Korea*
[3]*Linguistic and Cognitive Science, Hankuk University of Foreign Studies, Yongin, Korea*
[4]*Chinese Culture, Sogang University, Sincheon, Korea*
[5]*Human Centered Artificial Intelligence, Sangmyung University, Jongno, Korea*
[6]*Computer Engineering, Hankuk University of Foreign Studies, Yongin, Korea*

e2huh321@gmail.com[1], *bg000311@gmail.com*[2], *jisunny0242@gmail.com*[3], *cksgh0984@gmail.com*[4], *tlgus5627@gmail.com*[5], *so517273@gmail.com*[6]

Abstract

This paper aims to create a dataset and model for verifying language sensitivity with the goal of identifying non-sensitive words and enhancing awareness of language sensitivity. To build the models, we gathered corpus data from articles and documents in AI Hub and generated data using ChatGPT. Additionally, we constructed a sensitivity dictionary based on research papers on language sensitivity. Once the model is operational, it categorizes non-sensitive words and suggests suitable replacements. It also provides results for sentiment analysis, helping to differentiate between sentiment analysis and language sensitivity. We employed KcElectra for both sentiment analysis and language sensitivity models. This study represents the first convergence approach to language sensitivity and engineering, contributing to the improvement of sensitivity in the Korean language.

Keywords: *Language Sensitivity, Sentiment Analysis, Sensitivity Dictionary, Korean NLP, AI Technique, Validation of Language Sensitivity.*

1. Introduction

Language sensitivity means that when using language, you must sensitively consider various situations. Our society uses word expressions that contain discriminatory meanings, prejudices, and stereotypes with no intent. The letter '모' in the Korean word '유모차(stroller)' means mother. Therefore, the word '유모차' has the meaning of 'a carriage carried by a mother'. Then, Are fathers, grandparents carrying their children considered in that word? As in these cases, language sensitivity appears with suspicion in situations where someone is discriminated against or not considered within a word. As society changes, the importance of language

Advanced Engineering and ICT-Convergence Proceedings (AEICP)
ISSN : 2635-4586

sensitivity is increasing, and there is a growing trend of opinions that words with such discriminatory meanings should be replaced with other words. Therefore, we create a model that can distinguish between high and low language sensitivity by fine-tuning the KCELECTRA model.

2. Methodology

2.1. Dataset and Construction of Language Sensitivity Dictionary

2.1.1. Dataset

We collected 400,000 pieces of data by gathering AI Hub data (machine reading data for administrative documents, document summary text data, news reading data, thematic textual daily data, online colloquial corpus data, and broadcast content script summary data), newspaper data from National Institute of the Korean Language's Corpus of Everyone, sentences generated by ChatGPT, and posts and comments from internet communities (DCInside, Natepann, Theqoo, and Naver cafe). This allowed us to construct both written and spoken language datasets.

Since there was no existing data related to language sensitivity, we conducted manual labeling ourselves. Utilizing six categories named disability and medical history, gender and family, social status, origin, profanity, and others, we labeled which areas had low language sensitivity. Initially, we referred to domestic papers and publications to select words with low language sensitivity. Sentences containing these words were then labeled (Table 1). If two or more words in one sentence have low language sensitivity, multi-labeling was performed.

Table 1. Example of Labeling

Sentence	DM	GF	Social status	Origin	Profanity	Other	Sensitivity _LOW	Sensitivity _HIGH
나는 새 유모차를 샀어	0	1	0	0	0	0	1	0
불우이웃을 위해 반팔티를 기부했다	1	0	1	0	0	0	1	0

In addition, we collected additional data to rectify any possible misjudgments resulting from homophones. For instance, the phrase ‘편부 가정을 위한 자녀 돌봄 서비스’ contains ‘편부’, refers to alone father, and the letter ‘편’, meaning to lean, so it is a word with low language sensitivity. To avoid confusion, we collected additional examples, such as ‘1 편부터 3 편까지 모두 재미있어’, to ensure the model distinguishes between low language sensitivity and standard usage of similar sounding words.

2.1.2. Dictionary construction

Based on the 2019 National Human Rights Commission statistics 107p, we set up six categories of disability and medical history, gender and family, social status, origin, profanity, and other, and built a language sensitivity dictionary based on 20 articles.

The dictionary has a 1:1 correspondence between *from* and *to*, with expressions that do not consider language sensitivity in *from* and expressions that consider language sensitivity in *to*. At this time, if an expression that does not consider language sensitivity has a condescending or demeaning meaning based on the standard Korean dictionary, an alternative word is not recommended, and an alternative word is selected

according to papers and dictionary studies.

Table 2. Example of Dictionary [Gender and Family]

Index	From	To
ㄱ	가정부	가사 관리사
	결손가족	가족
	결손가정	가정
	김 여사	운전 미숙자
	김여사	운전 미숙자
	고아원	아동복지시설
	경력단절여성	고용중단여성
	과부	故 000 씨 배우자

The dictionary is updated in the following way. For sentences identified as low language sensitivity by the language sensitivity classification model, extract only nouns using the morphological analyzer, and compare the word combinations generated using the n-gram method with the dictionary's list of *from* words. If there is a matching word in the dictionary, it is output as a low language sensitivity expression, and if not, it is saved as a newly discovered low language sensitivity expression. The dictionary will then be regularly checked to identify new low-sensitivity words that have arisen since the time of the study, and the dictionary will be continuously expanded.

2.2. Sentiment Analysis Model

To effectively discern the differences in sentiment sensitivity model, we trained our model using data from Naver movie reviews and online shopping mall reviews. We fine-tuned the Kobert model, which was the most performant among the available pre-trained models, creating a classification model. The model achieved a high accuracy of 0.8947 and an F1 - score of 0.8902, indicating its robustness in sentiment-sensitive tasks.

2.3. Language Sensitivity Analysis Model

In this study, we utilized the KCELECTRA model, which had been pretrained and fine- tuned with a diverse set of textual materials including comments, news articles, and community posts, to assess language sensitivity. The model employed binary classification to determine whether a sentence was of low or high sensitivity to language. Subsequently, a hexary classification was used to identify the category of language sensitivity to which a sentence belonged. For instance, the sentence "나 오늘 반팔 티 입었어." was classified as having low language sensitivity and was determined to fall under the category of disability.

2.4. Alternative Expressions Proposal

In this study, we have developed an automated system that identifies sentiment-sensitive words within Korean sentences and suggests alternative expressions. Utilizing natural language processing technologies, the system evaluates the sentiment sensitivity of text and proposes alternatives that maintain contextual appropriateness while enhancing language use.

The system tokenizes input sentences through morphological analysis and filters out words of specific parts of speech based on a predefined set. Subsequently, it forms combinations of consecutive words using bigrams and trigram. Also regular expressions are employed to refine sentences, and the identified words are evaluated based on a predefined sentiment sensitivity index.

The proposed system effectively identified sentiment - sensitive words within sentences and was able to improve the overall quality of language use by suggesting appropriate alternatives. The performance of the

system was evaluated based on the accuracy of identifying sentiment - sensitive words and the contextual appropriateness of the suggested alternatives.

3. Result

3.1. Model Evaluation

In the realm of computational linguistics, our study has made strides by leveraging a language sensitivity model capable of discerning varying degrees of sensitivity within text. Through rigorous comparative analysis, we evaluated the performance of two prominent models: Kobert and KcELECTRA, based on a binary classification framework. Also when applied to a hexary classification task, the KcELECTRA model maintained its remarkable performance.

Table 3. Classification Model Evaluation Metrics

Evaluation Metrics	KOBERT for binary classification	KCELECTRA for binary classification	KCELECTRA for multi-class classification
Accuracy	0.9532	0.9852	0.9812
Precision	0.9643	0.9655	0.9433
Recall	0.9424	0.9865	0.9943
F1	0.9535	0.9763	0.9688

These results substantiate the efficacy of the KcELECTRA model in accurately classifying language sensitivity across multiple dimensions.

3.2. Output Results

Table 4. Language Sensitivity Analysis Model Result

Sentence	Disability	Gender	Social Status	Origin	Profanity	Other
관악농업협동조합은 독거노인 및 장애우 가정, 취약 계층을 대상으로 설날맞이 사랑의 쌀 전달식을 개최하였다.	0.7778	0.0001	0.9998	0.00002	0.00002	0.2292

As a result, "관악농업협동조합은 설날을 맞아 독거노인, 장애우 가정 및 사회적 취약 계층을 대상으로 사랑의 쌀 전달식을 개최하였다." targeted the elderly living alone, households with disabilities, and the socially vulnerable. When this sentence was processed through a language sensitivity category model, terms such as "elderly living alone," "people with disabilities," and "socially vulnerable" were identified as discriminatory towards disabilities and social status. The model demonstrates its effectiveness in discerning categories related to language sensitivity. Furthermore, by inserting new terms into the model, we can verify the language sensitivity of the terms and identify their associated categories.

3.3. Limitation

Our language sensitivity model has a critical limitation tied to its binary labeling system based on specific keywords, which oversimplifies the nuanced and context-sensitive nature of the subject. This reductionist approach may artificially inflate accuracy metrics, highlighting the challenge of applying machine learning to the subtleties of language sensitivity. Relying on keyword detection oversimplifies a complex concept, raising

questions about the model's broader applicability. Despite these challenges, our dataset compilation and engineering analysis represent an initial step toward quantitatively studying a traditionally qualitative field, paving the way for more sophisticated models to navigate the intricacies of language sensitivity.

4. Conclusion

The study focused on analyzing and validating modern languages based on societal changes and language sensitivity across six categories. It introduced models for language and emotion analysis, highlighting differences between the two. These models aid in improving language-aware artificial intelligence by filtering out less sensitive words, promoting better language use. They can also help companies and institutions review marketing or official content for higher sensitivity, potentially reducing costs and fostering stronger consumer connections. Ultimately, this study pioneers an engineering approach towards language sensitivity, integrating humanities elements, and lays a foundation for future research.

Acknowledgement

Acknowledgement title is not numbered. Type the acknowledgement in this format 'This work was supported by (institution to acknowledge) in (year).'

Note: Manuscripts in which references are not in this format will be returned without review.

References

[1] National Human Rights Commission of Korea. (2019). Statistics of the National Human Rights Commission of Korea.

[2] Lee, Y. J., Hwang, J. S., Kang, H. J., Kang, H. J., Kim, B. M., Kim, E. H., Jeon, J. H., Lee, S. H., Shin, H. S., Shin, E. K., & Jung, S. H. (2019). Cultural diversity supporters [Malmoe] activity materials collection. Gimhae Cultural Foundation.

[3] Lee, J. B. (2013). Regional discriminatory expressions on social networking services (SNS). 어문학 [Linguistic Studies], 120, 55-83.

[4] Park, M. S., & Choo, J. H. (2017). The current status and countermeasures of hate speech. 형사정책연구원 연구총서 [Korean Institute of Criminology Research Series], 1-373.

[5] Seo, H. J. (2023). Compilation and usage of discriminatory and authoritative expressions for public document evaluation. 공공언어학 [Public Language Studies], 9(0), 93-138.

[6] Shim, H. R., & Leeum Team. (2021). Workbook for finding regional discriminatory language. Hope Factory.

[7] Yoon, J. G. (2020). A collection of words that harm cultural diversity [Malmoe 2]. Gimhae Cultural Foundation.

[8] Park, E. H. (2019). A corpus-based study on the use and perception of occupation-discriminatory language. 사회언어학 [Journal of Sociolinguistics], 27(4), 89-116.