K-Datasets: The More The Merrier?

Chaewon Yun
Max Planck Institute for Human Development
yun@mpib-berlin.mpg.de

Joonha Jeon
AI Ethics Newsletter Korea
realjoonha@gmail.com

Koh Achim
AI Ethics Newsletter Korea
achim.koh@gmail.com

Linguistic diversity in training datasets is crucial for AI development to be globally inclusive. In this provocation, we address practical challenges of developing culturally-rich datasets based on a comprehensive literature review of Korean datasets. We especially focus on datasets on bias, hate speech, and abusive language, as their context-dependency presents greater pertinence to cultural inclusion. This case study provides implications on globally inclusive AI culture by investigating the specificities of Korean datasets, highlighting pitfalls and challenges relevant to similar tasks in lower-resource languages[1][2].

We find that one of the common pitfalls in Korean bias datasets is the lack of diversity in data sources; collecting local data does not automatically guarantee the development of a culturally representative dataset. Literature review shows that about half of the datasets use comments from online news portals, such as NAVER's[1], which collect and republish news articles. This reflects the unique characteristic of the Korean internet where about 70 percent of users rely on them for news consumption[3]. However, populations leaving comments on such news articles are not representative of Korean culture[2], and nor are online communities such as DCinside, another commonly used source[3][4].

Another practical challenge is the problem of limited resources, which can intensify the vulnerability of the datasets in terms of bias. As dataset research increasingly relies on language models to generate and evaluate data to fight data scarcity, encoded biases could be exacerbated and leave inclusiveness behind[5]. Limited resources can also introduce methodological pitfalls such as very few annotators or lack of diversity in groups, as is the case for some works we surveyed[4], risking aggravated annotator bias and reduced quality.[5]

Most of the research we reviewed motivates their work on the incompatibility of English datasets in Korean context due to the cultural differences. However, operationalization of concepts like bias, hate, or even social compatibility, when uninformed by existing work in relevant fields, can lead to questionable consequences. While different definitions and categorizations adopted by the datasets provide a lens to the diverse specificities of local context, it is a prerequisite to operationalize such social constructs with scientifically robust methods, through careful curating, processing, and application of such concepts in practice.[6]

---

[1] https://news.naver.com/
[2] Statistics show about 10% of users account for 70% of comments and the discrepancy between genders are significant. (https://datalab.naver.com/commentStat/news.naver)
[3] https://www.dcinside.com; Similar limitations have been criticized in English datasets regarding extensive use of Reddit data which over-represents certain populations[5].
[4] The full list of papers surveyed can be found in the Appendix.
[5] Yet another question is whether we can duly represent cultural diversity through crowd-sourced quantitative approaches[6].
[6] In addition, a substantial portion of the datasets are affiliated with Korean companies that run internet platforms or services. Hence, the motivation for developing such datasets is connected to the interest of the company's operation, rather than representing Korean culture or the Korean speaking population that is outside the scope of their user bases.

Despite such complications, it is imperative to put an effort towards more and diverse cultural representations in AI. Our analysis of Korean bias datasets show that adding one language to the list is not enough for inclusive global AI culture. To represent complex social phenomena such as hate speech, participatory design[7,8] and dataset audit[9,10,11] for encoding values to datasets have been suggested to overcome the limitation of quantitative approach. By reviewing Korean datasets from a critical perspective, we argue that making inclusive, rich, and diverse AI culture can be achieved by engaging in multi-disciplinary discourses that transcend the boundary of computational or technical solutions.

References

[1] Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
[2] Singh, S., Vargus, F., Dsouza, D., Karlsson, B. F., Mahendiran, A., Ko, W. Y., ... & Hooker, S. (2024). Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.
[3] Korea Press Foundation (2023). Media Users in Korea 2023. https://www.kpf.or.kr/synap/skin/doc.html?fn=1706080724689.pdf&rs=/synap/result/research/
[4] J, Kwon. (2011) From masculinity to cybermasculinity: Marginalizing the other in “DCinside”, In McCarthy, C. R., Greenhalgh-Spencer, H., & Mejia, R. (Eds.) (2011). *New Times: Making Sense of Critical/Cultural Theory in a Digital Age*. Peter Lang Publishing.
[5] Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.
[6] Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power?. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP), 1-14.
[7] Pérez-Escolar, M., & Noguera-Vivo, J. M. (2022). *Hate speech and polarization in participatory society*. Taylor & Francis.
[8] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, et al.. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.
[9] Birhane, A., Han, S., Boddeti, V., & Luccioni, S. (2024). Into the LAION’s Den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems*, 36.
[10] Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-29.
[11] DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022, April). Toward User-Driven Algorithm Auditing: Investigating users’ strategies for uncovering harmful

algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Appendix

Dataset	**Source**	**Categories**	**Annotators**	**Industry Affiliation**
KoBBQ [1]	Cultural adaptation of BBQ dataset, author-generated templates for Korean culture specific cases	Age, Educational Background, Political Orientation, Family Structure, Domestic Area of Origin, Sexual Orientation, Socio-Economic Status, Religion, Race/Ethnicity/Nationality, Physical Appearance, Gender Identity, Disability Status	100 Korean individuals per survey question (Annotator information not available)	NAVER
K-OMG [2]	LLM-generated	HumanOrAI, Relevance, Offensiveness, and Fluency	5 native speakers of Korean	
KoTox [3]	LLM-generated	Unethical instruction-output pairs - response annotation: ethical, neutral, unethical, irrelevant, incorrect	Annotator information not available	
SQuARe [4]	LLM-generated	Sensitive questions(contentious, ethical, predictive), Acceptable (inclusive with social groups, inclusive with opinions, ethically aware, nonpredictive, objective, indirect)	258 crowd workers (Annotation Demographics included)	NAVER
KoSBi [5]	LLM-generated	Biased (Stereotypes, Prejudice, Discrimination), Other	200 crowd workers (Annotation Demographics included)	NAVER
UnSmile [6]	Comments from news section in NAVER and Daum, Online community websites (DC Inside, Ilbe, Womad, and Today Humor)	Race and Nationality, Religion, Regionalism, Ageism, Misogyny, Sexual Minorities, and Male	13 annotators (researchers with a master's degree or higher in social science including 7 authors)	Smilegate
BEEP! [7]	Comments from the Korean entertainment news aggregation platform	Social bias (Gender, others, none), Hate speech (hate, offensive, none)	32 annotators (29 workers from a crowdsourcing platform DeepNatural AI7 and three natural language processing (NLP) researchers)	
K-HATERS [8]	Comments from the news section in NAVER (Society, World news, Politics sections) , BEEP! dataset (entertainment section)	Target-specific (Group: gender, age, race, religion, politics, job, disability, Individual, Other) and Fine-grained ratings (Insult, Swear words, Obscenity, Threat)	405 annotators (Annotator information not available)	SelectStar
KOLD [9]	Titles and comments from NAVER news articles and YouTube videos	Target group (Gender & Sexual Orientation, Race, Ethnicity & Nationality, Political Affiliation, Religion, Miscellaneous)	3124 annotators (Annotator information not available)	NAVER, SoftlyAI
K-MHaS [10]	Existing dataset of NAVER news comments and BEEP! dataset	Binary classification ('Hate Speech' or 'Not Hate Speech'), Fine-grained classification (8 labels - politics, origin, physical, age, gender, religion, race, profanity) or 'Not Hate Speech"	4 native speakers of Korean	
KODOLI [11]	Comments from news section in NAVER, online Korean communities, such as Dcinside,, and existing dataset of texts of NAVER shopping and Steam	Binary classification (Abuse, non-abuse)	11 annotators (57% men, 43% women, undergraduate and graduate students)	
APEACH [12]	Crowd workers-generated by using pseudo classifier which used texts of an online community YourSSU and news comments	Binary classification (Hate speech or not), Topics (Racism, sexual harassment, gender stereotypes, eating habits, appearance, age and social status, education, origin and residence, disabled, nationality)	154 crowd workers (Annotator information not available)	Kakao
KoMultiText [13]	Online community Dcinside ("Real-time Best Gallery")	Preferences, Profanities, Nine types of Bias (Gender, politics, nation, race, region, generation, social hierarchy, appearance, others)	4 annotators	
KOAS [14]	Youtube, Dcinside, and existing dataset of NAVER Movie Review	Binary classification (Abuse, non-abuse)	3 annotators	

Table 1: Literature Review on Korean Bias, Hate Speech, Abusive Language, and Social Acceptability Datasets

[1] Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., & Lee, H. (2023). Kobbq: Korean bias benchmark for question answering. *arXiv preprint arXiv:2307.16778*.
[2] Shin, J., Song, H., Lee, H., Gaim, F., & Park, J. C. (2023, November). Generation of Korean Offensive Language by Leveraging Large Language Models via Prompt Design. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 960-979).
[3] Byun, S., Jang, D., Jo, H., & Shin, H. (2023, November). Automatic Construction of a Korean Toxic Instruction Dataset for Ethical Tuning of Large Language Models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
[4] Lee, H., Hong, S., Park, J., Kim, T., Cha, M., Choi, Y., ... & Ha, J. W. (2023). SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created Through Human-Machine Collaboration. *arXiv preprint arXiv:2305.17696*.
[5] Lee, H., Hong, S., Park, J., Kim, T., Kim, G., & Ha, J. W. (2023). KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Application. *arXiv preprint arXiv:2305.17701*.
[6] Kang, T., Kwon, E., Lee, J., Nam, Y., Song, J., & Suh, J. (2022). Korean Online Hate Speech Dataset for Multilabel Classification: How Can Social Science Improve Dataset on Hate Speech?. *arXiv preprint arXiv:2204.03262*.
[7] Moon, J., Cho, W. I., & Lee, J. (2020). BEEP! Korean corpus of online news comments for toxic speech detection. *arXiv preprint arXiv:2005.12503*.
[8] Park, C., Kim, S., Park, K., & Park, K. (2023). K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings. *arXiv preprint arXiv:2310.15439*.
[9] Jeong, Y., Oh, J., Ahn, J., Lee, J., Moon, J., Park, S., & Oh, A. (2022). KOLD: korean offensive language dataset. *arXiv preprint arXiv:2205.11315*.
[10] 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*.
[11] Park, S. H., Kim, K. M., Lee, O. J., Kang, Y., Lee, J., Lee, S. M., & Lee, S. (2023, May). “Why do I feel offended?”-Korean Dataset for Offensive Language Identification. In *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 1112-1123).
[12] Yang, K., Jang, W., & Cho, W. I. (2022). APEACH: Attacking Pejorative Expressions with Analysis on Crowd-Generated Hate Speech Evaluation Datasets. *arXiv preprint arXiv:2202.12459*.
[13] Choi, D., Song, J., Lee, E., Park, H., & Na, D. (2023, November). KoMultiText: Large-Scale Korean Text Dataset for Classifying Biased Speech in Real-World Online Services. In *Socially Responsible Language Modelling Research*.
[14] Park, S. H., Kim, K. M., Cho, S., Park, J. H., Park, H., Kim, H., ... & Lee, S. (2021, November). KOAS: Korean text offensiveness analysis system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 72-78).