

BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection

Jihyung Moon^{*,†,1}, Won Ik Cho^{*,2}, Junbum Lee³

Department of Industrial Engineering¹,
Department of Electrical and Computer Engineering and INMC²,
Graduate School of Data Science³,
Seoul National University, Seoul
{ans1107, tsatsuki, beomi}@snu.ac.kr

Abstract

Toxic comments in online platforms are an unavoidable social issue under the cloak of anonymity. Hate speech detection has been actively done for languages such as English, German, or Italian, where manually labeled corpus has been released. In this work, we first present 9.4K manually labeled entertainment news comments for identifying Korean toxic speech, collected from a widely used online news platform in Korea. The comments are annotated regarding social bias and hate speech since both aspects are correlated. The inter-annotator agreement Krippendorff's alpha score is 0.492 and 0.496, respectively. We provide benchmarks using CharCNN, BiLSTM, and BERT, where BERT achieves the highest score on all tasks. The models generally display better performance on bias identification, since the hate speech detection is a more subjective issue. Additionally, when BERT is trained with bias label for hate speech detection, the prediction score increases, implying that bias and hate are intertwined. We make our dataset publicly available and open competitions with the corpus and benchmarks.

1 Introduction

Online anonymity provides freedom of speech to many people and lets them speak their opinions in public. However, anonymous speech also has a negative impact on society and individuals (Banks, 2010). With anonymity safeguards, individuals easily express hatred against others based on their superficial characteristics such as gender, sexual orientation, and age (ElSherief et al., 2018). Sometimes the hostility leaks to the well-known people who are considered to be the representatives of targeted attributes.

Recently, Korea had suffered a series of tragic incidents of two young celebrities that are presumed to be caused by toxic comments (Fortin, 2019; McCurry, 2019a,b). Since the incidents, two major web portals in Korea decided to close the comment system in their entertainment news aggregating service (Yeo, 2019; Yim, 2020). Even though the toxic comments are now avoidable in those platforms, the fundamental problem has not been solved yet.

To cope with the social issue, we propose the first Korean corpus annotated for toxic speech detection. Specifically, our dataset consists of 9.4K comments from Korean online entertainment news articles. Each comment is annotated on two aspects, the existence of social bias and hate speech, given that hate speech is closely related to bias (Boeckmann and Turpin-Petrosino, 2002; Waseem and Hovy, 2016; Davidson et al., 2017). Considering the context of Korean entertainment news where public figures encounter stereotypes mostly intertwined with gender, we weigh more on the prevalent bias. For hate speech, our label categorization refers that of Davidson et al. (2017), namely *hate*, *offensive*, and *none*.

The main contributions of this work are as follows:

- We release the first Korean corpus manually annotated on two major toxic attributes, namely bias and hate¹.
- We hold Kaggle competitions^{2,3} and provide benchmarks to boost further research development.
- We observe that in our study, hate speech detection benefits the additional bias context.

^{*}Both authors contributed equally to this manuscript.

[†]This work was done after the graduation.

¹<https://github.com/kocohub/korean-hate-speech>

²www.kaggle.com/c/korean-gender-bias-detection

³www.kaggle.com/c/korean-bias-detection

⁴www.kaggle.com/c/korean-hate-speech-detection

2 Related Work

The construction of hate speech corpus has been explored for a limited number of languages, such as English (Waseem and Hovy, 2016; Davidson et al., 2017; Zampieri et al., 2019; Basile et al., 2019), Spanish (Basile et al., 2019), Polish (Ptaszynski et al., 2019), Portuguese (Fortuna et al., 2019), and Italian (Sanguinetti et al., 2018).

For Korean, works on abusive language have mainly focused on the qualitative discussion of the terminology (Hong, 2016), whereas reliable and manual annotation of the corpus has not yet been undertaken. Though profanity termbases are currently available⁵⁶, term matching approach frequently makes false predictions (e.g., neologism, polysemy, use-mention distinction), and more importantly, not all hate speech are detectable using such terms (Zhang et al., 2018).

In addition, hate speech is situated within the context of social bias (Boeckmann and Turpin-Petrosino, 2002). Waseem and Hovy (2016) and Davidson et al. (2017) attended to bias in terms of hate speech, however, their interest was mainly in texts that explicitly exhibit sexist or racist terms. In this paper, we consider both explicit and implicit stereotypes, and scrutinize how these are related to hate speech.

3 Collection

We constructed the Korean hate speech corpus using the comments from a popular domestic entertainment news aggregation platform. Users had been able to leave comments on each article before the recent overhaul (Yim, 2020), and we had scrapped the comments from the most-viewed articles.

In total, we retrieved 10,403,368 comments from 23,700 articles published from January 1, 2018 to February 29, 2020. We draw 1,580 articles using stratified sampling and extract the top 20 comments ranked in the order of Wilson score (Wilson, 1927) on the downvote for each article. Then, we remove duplicate comments, single token comments (to eliminate ambiguous ones), and comments composed with more than 100 characters (that could convey various opinions). Finally, 10K comments are randomly selected among the rest for annotation.

⁵<https://github.com/doublems/korean-bad-words>

⁶<https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words>

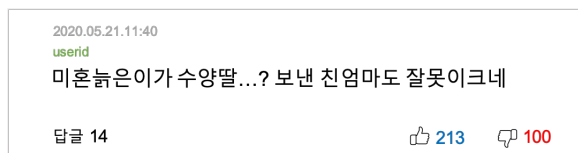


Figure 1: A sample comment from the online news platform. It is composed of six parts: written date and time, masked user id, content, the number of replies, and the number of up/down votes (from top left to bottom right).

We prepared other 2M comments by gathering the top 100 sorted with the same score for all articles and removed with any overlaps regarding the above 10K comments. This additional corpus is distributed without labels, expected to be useful for pre-training language models on Korean online text.

4 Annotation

The annotation was performed by 32 annotators consisting of 29 workers from a crowdsourcing platform *DeepNatural AI*⁷ and three natural language processing (NLP) researchers. Every comment was provided to three random annotators to assign the majority decision. Annotators are asked to answer two three-choice questions for each comment:

1. What kind of bias does the comment contain?
 - *Gender bias, Other biases, or None*
2. Which is the adequate category for the comment in terms of hate speech?
 - *Hate, Offensive, or None*

They are allowed to skip comments which are too ambiguous to decide. Detailed instructions are described in Appendix A. Note that this is the first guideline of social bias and hate speech on Korean online comments.

4.1 Social Bias

Since hate speech is situated within the context of social bias (Boeckmann and Turpin-Petrosino, 2002), we first identify the bias implicated in the comment. Social bias is defined as a preconceived evaluation or prejudice towards a person/group with certain social characteristics: gender, political affiliation, religion, beauty, age, disability, race, or others. Although our main interest is on gender bias, other issues are not to be underestimated.

⁷<https://app.deepnatural.ai/>

Thus, we separate bias labels into three: whether the given text contains gender-related bias, other biases, or none of them. Additionally, we introduce a binary version of the corpus, which counts only the gender bias, that is prevalent among the entertainment news comments.

The inter-annotator agreement (IAA) of the label is calculated based on Krippendorff’s alpha (Krippendorff, 2011) that takes into account an arbitrary number of annotators labeling any number of instances. IAA for the ternary classes is 0.492, which means that the agreement is moderate. For the binary case, we obtained 0.767, which implies that the identification of gender and sexuality-related bias reaches quite a substantial agreement.

4.2 Hate Speech

Hate speech is difficult to be identified, especially for the comments which are context-sensitive. Since annotators are not given additional information, labeling would be diversified due to the difference in pragmatic intuition and background knowledge thereof. To collect reliable hate speech annotation, we attempt to establish a precise and clear guideline.

We consider three categories for hate speech: *hate*, *offensive but not hate*, and *none*. As socially agreed definition lacks for Korean⁸, we refer to the hate speech policies of [Youtube](#); [Facebook](#); [Twitter](#). Drawing upon those, we define hate speech in our study as follows:

- If a comment explicitly expresses hatred against individual/group based on any of the following attributes: sex, gender, sexual orientation, gender identity, age, appearance, social status, religious affiliation, military service, disease or disability, ethnicity, and national origin
- If a comment severely insults or attacks individual/group; this includes sexual harassment, humiliation, and derogation

However, note that not all the rude or aggressive comments necessarily belong to the above definition, as argued in [Davidson et al. \(2017\)](#). We often see comments that are offensive to certain individuals/groups in a qualitatively different manner. We identify these as offensive and set the boundary as follows:

⁸Though a government report is available for the Korean language ([Hong, 2016](#)), we could not reach a fine extension to the quantitative study on online spaces.

(%)	Hate	Offensive	None	Sum (Bias)
Gender	10.15	4.58	0.98	15.71
Others	7.48	8.94	1.74	18.16
None	7.48	19.13	39.08	65.70
Sum (Hate)	25.11	32.66	41.80	100.00

Table 1: Distribution of the annotated corpus.

- If a comment conveys sarcasm via rhetorical expression or irony
- If a comment states an opinion in an unethical, rude, coarse, or uncivilized manner
- If a comment implicitly attacks individual/group while leaving rooms to be considered as freedom of speech

The instances that do not meet the boundaries above were categorized as *none*. The IAA on the hate categories is $\alpha = 0.496$, which implies a moderate agreement.

5 Corpus

Release From the 10k manually annotated corpus, we discard 659 instances that are either skipped or failed to reach an agreement. We split the final dataset into the train (7,896), validation (471), and test set (974) and released it on the Kaggle platform to leverage the leaderboard system. For a fair competition, labels on the test set are not disclosed. Titles of source articles for each comment are also provided, to help participants exploit context information.

Class distribution Table 1 depicts how the classes are composed of. The bias category distribution in our corpus is skewed towards *none*, while that of *hate* category is quite balanced. We also confirm that the existence of hate speech is correlated with the existence of social bias. In other words, when a comment incorporates a social bias, it is likely to contain hate or offensive speech.

6 Benchmark Experiment

6.1 Models

We implemented three baseline classifiers: character-level convolutional neural network (CharCNN) ([Zhang et al., 2015](#)), bidirectional long short-term memory (BiLSTM) ([Schuster and Paliwal, 1997](#)), and bidirectional encoder representations from Transformer (BERT) ([Devlin et al., 2018](#)) based model. For BERT, we adopt

True categories	Hate	20.23	7.7	0.31
	Offensive	16.02	20.02	2.57
	None	4.62	17.45	11.09
		Hate	Offensive	None
		Predicted categories		

(a) BERT predictions

True categories	Hate	21.36	5.75	1.13
	Offensive	17.25	14.68	6.67
	None	3.39	9.24	20.53
		Hate	Offensive	None
		Predicted categories		

(b) BERT predictions with bias label

Figure 2: Confusion matrix on the model inference of hate categories.

F1	Bias (binary)	Bias (ternary)	Hate
Term Matching	-	-	0.195
CharCNN	0.547	0.535	0.415
BiLSTM	0.302	0.291	0.340
BERT	0.681	0.633	0.525
BERT (+ bias)	-	-	0.569

Table 2: F1 score of benchmarks on the test set. Note that the term matching model checks the presence of hate or offensiveness. Therefore, in this case, we combine *hate* and *offensive* into a single category, turning the original ternary task into binary.

KoBERT⁹, a pre-trained module for the Korean language, and apply its tokenizer to BiLSTM as well. The detailed configurations are provided in Appendix B, and we additionally report the term matching approach using the aforementioned profanity terms to compare with the benchmarks.

6.2 Results

Table 2 depicts F1 score of the three baselines and the term matching model. The results demonstrate that the models trained on our corpus have an advantage over the term matching method. Compared with the benchmarks, BERT achieves the best performance for all the three tasks: binary and ternary bias identification tasks, and hate speech detection. Each model not only shows different performances but also presents different characteristics.

Bias detection When it comes to the gender-bias detection, the task benefits more on CharCNN than BiLSTM since the bias label is highly correlated with frequent gender terms (e.g., *he*, *she*, *man*, *woman*, ...) in the dataset. It is known that Char-

F1	Gender	Others	None	Bias (ternary)
CharCNN	0.519	0.259	0.826	0.535
BiLSTM	0.055	0.000	0.819	0.291
BERT	0.693	0.326	0.880	0.633

Table 3: Detailed results on macro-F1 of Bias (ternary)

CNN well captures the lexical components that are present in the document.

However, owing to that nature, CharCNN sometimes yields results that are overly influenced by the specific terms which cause false predictions. For example, the model fails to detect bias in “*What a long life for a GAY*” but guesses “*I think she is the prettiest among all the celebs*” to contain bias. CharCNN overlooks *GAY* while giving a wrong clue due to the existence of female pronouns, namely *she* in the latter.

Similar to the binary prediction task, CharCNN outperforms BiLSTM on ternary classification. Table 3 demonstrates that BiLSTM hardly identifies *gender* and *other* biases.

BERT detects both biases better than the other models. From the highest score obtained by BERT, we found that rich linguistic knowledge and semantic information is helpful for bias recognition.

We also observed that all the three models barely perform well on *others* (Table 3). To make up a system that covers the broad definition of *other* bias, it would be better to predict the label as the non-*gender* bias. For instance, it can be performed as a two-step prediction: the first step to distinguish whether the comment is biased or not and the second step to determine whether the biased comment is gender-related or not.

⁹<https://github.com/SKTBrain/KoBERT>

Hate speech detection For hate speech detection, all models faced performance degradation compared to the bias classification task, since the task is more challenging. Nonetheless, BERT is still the most successful, and we conjecture that hate speech detection also utilizes high-level semantic features. The significant performance gap between term matching and BERT explains how much our approach compensates for the false predictions mentioned in Section 2.

Provided *bias* label prepend to each comment as a special token, BERT exhibits better performance. As illustrated in Figure 2, additional bias context helps the model to distinguish *offensive* and *none* clearly. This implies our observation on the correlation between bias and hate is empirically supported.

7 Conclusions

In this data paper, we provide an annotated corpus that can be practically used for analysis and modeling on Korean toxic language, including hate speech and social bias. In specific, we construct a corpus of a total of 9.4K comments from online entertainment news service.

Our dataset has been made publicly accessible with baseline models. We launch Kaggle competitions using the corpus, which may facilitate the studies on toxic speech and ameliorate the cyberbullying issues. We hope our initial efforts can be supportive not only to NLP for social good, but also as a useful resource for discerning implicit bias and hate in online languages.

Acknowledgments

We greatly thank Hyunjoong Kim for providing financial support and Sangwoong Yoon for giving helpful comments.

References

James Banks. 2010. Regulating hate speech online. *International Review of Law, Computers & Technology*, 24(3):233–239.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Robert J Boeckmann and Carolyn Turpin-Petrosino. 2002. Understanding the harm of hate crime. *Journal of social issues*, 58(2):207–225.

Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media*.

Facebook. Facebook’s policy on hate speech. https://www.facebook.com/communitystandards/hate_speech. Accessed: 2020-04-19.

Jacey Fortin. 2019. [Sulli, south korean k-pop star and actress, is found dead](#). *New York Times*.

Paula Fortuna, João Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Seong Soo Hong. 2016. *Hate speech: Survey and Regulations*. National Human Rights Commission of the Republic of Korea.

Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Justin McCurry. 2019a. [K-pop singer goo hara found dead aged 28](#). *The Guardian*.

Justin McCurry. 2019b. [K-pop under scrutiny over ‘toxic fandom’ after death of sulli](#). *The Guardian*.

Michal Ptaszynski, Agata Pieciukiewicz, and Pawel Dybała. 2019. Results of the poleval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in polish twitter. *Proceedings of the PolEval2019 Workshop*, page 89.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Twitter. Twitter’s policy on hate speech. <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>. Accessed: 2020-04-19.

Zeeraq Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Junsuk Yeo. 2019. Kakao suspends online comments for entertainment articles after sulli’s death. *The Korea Herald*.

Hyunsu Yim. 2020. Why naver is finally shutting down comments on celebrity news. *The Korea Herald*.

Youtube. Youtube’s policy on hate speech. <https://support.google.com/youtube/answer/2801939?hl=en>. Accessed: 2020-04-19.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

A Annotation Guideline

A.1 Existence of social bias

The first property is to note which social bias is implicated in the comment. Here, social bias means hasty guess or prejudice that ‘a person/group with a certain social identity will display a certain characteristic or act in a biased way’. The three labels of the question are as follows.

1. Is there a gender-related bias, either explicit or implicit, in the text?
 - If the text includes bias for gender role, sexual orientation, sexual identity, and any thoughts on gender-related acts (e.g., “Wife must be obedient to her husband’s words”, or “Homosexual person will be prone to disease.”)
2. Are there any other kinds of bias in the text?
 - Other kinds of factors that are considered not gender-related but social bias, including race, background, nationality, ethnic group, political stance, skin color, religion, handicaps, age, appearance, richness, occupations, the absence of military service experience¹⁰, etc.
3. A comment that does not incorporate the bias

A.2 Amount of hate, insulting, or offense

The second property is how aggressive the comment is. Since the level of “aggressiveness” depends on the linguistic intuition of annotators, we set the following categorization to draw a borderline as precise as possible.

1. Is strong hate or insulting towards the article’s target or related figures, writers of the article or comments, etc. displayed in a comment?
 - In the case of insulting, it encompasses an expression that can severely harm the social status of the recipient.
 - In the case of hate, it is defined as an expression that displays aggressive stances towards individuals/groups with certain characteristics (gender role, sexual orientation, sexual identity, any thoughts on gender-related acts, race, background, nationality, ethnic group, political stance, skin color, religion, handicaps, age, appearance, richness, occupations, the absence of military service experience, etc.).
 - Additionally, it can include sexual harassment, notification of offensive rumors or facts, and coined terms for bad purposes or in bad use, etc.
 - Just an existence of bad words in the document does not always fall into this category.

¹⁰Frequently observable in Korea, where the military service is mandatory for males.

2. Although a comment is not as much hateful or insulting as the above, does it make the target or the reader feel offended?
 - It may contain rude or aggressive contents, such as bad words, though not to the extent of hate or insult.
 - It can emit sarcasm through rhetorical questions or irony.
 - It may encompass an unethical expression (e.g., jokes or irrelevant questions regarding the figures who passed away).
 - A comment conveying unidentified rumors can belong to this category.
3. A comment that does not incorporate any hatred or insulting

B Model Configuration

Note that each model’s configuration is the same for all tasks except for the last layer.

B.1 CharCNN

For character-level CNN, no specific tokenization was utilized. The sequence of Hangul characters was fed into the model at a maximum length of 150. The total number of characters was 1,685, including ‘[UNK]’ and ‘[PAD]’ token, and the embedding size was set to 300. 10 kernels were used, each with the size of [3,4,5]. At the final pooling layer, we used a fully connected network (FCN) of size 1,140, with a 0.5 dropout rate (Srivastava et al., 2014). The training was done for 6 epochs.

B.2 BiLSTM

For bidirectional LSTM, we had a vocab size of 4,322, with a maximum length of 256. We used BERT SentencePiece tokenizer (Kudo and Richardson, 2018). The width of the hidden layers was 512 ($=256 \times 2$), with four stacked layers. The dropout rate was set to 0.3. An FCN of size 1,024 was appended to the BiLSTM output to yield the final softmax layer. We trained the model for 15 epochs.

B.3 BERT

For BERT, a built-in SentencePiece tokenizer of KoBERT was adopted, which was also used for BiLSTM. We set a maximum length at 256 and ran the model for 10 epochs.