KOAS: Korean Text Offensiveness Analysis System

San-Hee Park[1*] **Kang-Min Kim**[3*] **Seonhee Cho**[1*] **Jun-Hyung Park**[1]
Hyuntae Park[2] **Hyuna Kim**[1] **Seongwon Chung**[1] **SangKeun Lee**[1,2]
[1] Department of Computer Science and Engineering [2] Department of Artificial Intelligence
Korea University, Seoul, Republic of Korea
[3] Department of Data Science, The Catholic University of Korea, Bucheon, Republic of Korea
carpediem20@korea.ac.kr kangmin89@catholic.ac.kr
{ehcho8564, irish07, pht0639}@korea.ac.kr
{kiipo0623, syc1013, yalphy}@korea.ac.kr

Abstract

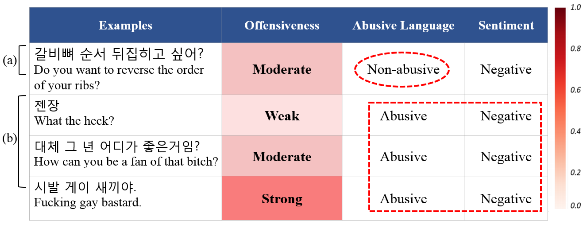
***Warning**: This manuscript contains a certain level of offensive expression.*

As communication through social media platforms has grown immensely, the increasing prevalence of offensive language online has become a critical problem. Notably in Korea, one of the countries with the highest Internet usage, automatic detection of offensive expressions has recently been brought to attention. However, morphological richness and complex syntax of Korean causes difficulties in neural model training. Furthermore, most of previous studies mainly focus on the detection of abusive language, disregarding implicit offensiveness and underestimating a different degree of intensity. To tackle these problems, we present KOAS, a system that fully exploits both contextual and linguistic features and estimates an offensiveness score for a text. We carefully designed KOAS with a multi-task learning framework and constructed a Korean dataset for offensive analysis from various domains. Refer for a detailed demonstration. [1]

1 Introduction

Online communities and social media have become the mainstream platforms of communication. This has also led to unwanted developments – an increasing use of offensive language through online platforms. Consequently, analyzing texts and detecting offensive expressions has become a critical issue (Nobata et al., 2016). However, manual detection of offensive texts is infeasible owing to the increasing popularity of social networks (Kennedy et al., 2017). Notably in South Korea, high internet accessibility and social media usage[2] have stimulated a dire need for a system that analyzes Korean text and its offensiveness (Moon et al., 2020a).

Despite the recent success of offensive language detection on English text (Mishra et al., 2019), handling Korean texts is quite challenging. Owing to the high-context cultural characteristics of Korean language culture (Merkin, 2009), Korean offensive expressions tend to be expressed in a subtle and figurative way without explicit abusive expressions as illustrated in Figure 1(a). Additionally, large vocabulary and the complex syntax as a morphologically rich and agglutinative language (Song, 2006) often hinders the model's learning (Kim et al., 2018; Passban et al., 2018).



	Examples	Offensiveness	Abusive Language	Sentiment
(a)	갈비뼈 순서 뒤집히고 싶어? Do you want to reverse the order of your ribs?	**Moderate**	Non-abusive	Negative
(b)	젠장 What the heck?	**Weak**	Abusive	Negative
	대체 그 년 어디가 좋은거임? How can you be a fan of that bitch?	**Moderate**	Abusive	Negative
	시발 게이 새끼야. Fucking gay bastard.	**Strong**	Abusive	Negative

1.0
0.8
0.6
0.4
0.2
0.0

Figure 1: (a) Illustration of an example of offensive sentence without any explicit profanity, which would be classified as "non-abusive" in abusive language detection. (b) Illustration of sentences with different intensity of offensiveness that is not distinguished properly in the discrete classification tasks.

Another substantial problem in text analysis is that the intensity of offensiveness in text is often neglected. As shown in Figure 1(b), the sentence may address the different intensity depending on the degree of frequency and explicitness of the offensive expression (Jay and Janschewitz, 2008; Jay, 2009). However, most researches focus on simply classifying sentences into discrete, sometimes binary, categories (Kennedy et al., 2017; Patwa et al., 2020; Mishra et al., 2019) and treat sentences with different intensity as the same type, *"negative"*, or *"abusive"* for instance.

In this demo, we present KOAS, a system that estimates the score of offensiveness in Korean text. Since the degree of offensiveness is different from of abusive detection or sentiment analysis, we develop a scoring function for offensiveness. The score is to quantify how much negative feelings each sentence can cause to readers, or how offensively it can be read. To this end, KOAS internally conducts two classification tasks, abusive language detection and sentiment analysis. An offensiveness score is then elicited from the outputs of two internal tasks. While computing the score, KOAS inte-

[*] These authors contributed equally to this work.
[1] https://www.youtube.com/watch?v=xtQv7GKOaeg (The KOAS link will be updated later.)
[2] https://datareportal.com/reports/digital-2020-south-korea

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 72–78
November 7–11, 2021. ©2021 Association for Computational Linguistics

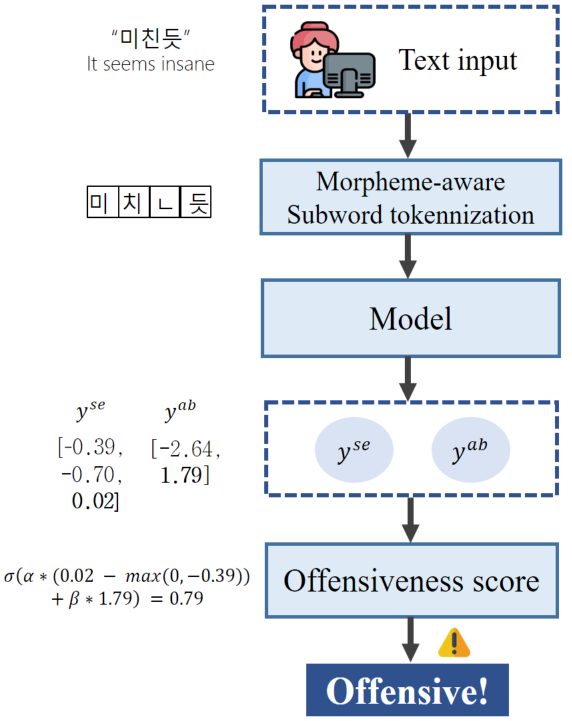


Figure 2: This is a flowchart of KOAS, from taking a Korean text input to eliciting the offensiveness score. y^{se} and y^{ab} denote output vectors of the sentiment analysis and abusiveness detection respectively.

grates the semantic perspective where it detects explicit abusive expressions and semantic perspective where it captures implicit nuance and tone of the text.

Since the notion of continuous degree of offensiveness has not been researched in depth yet, there are no appropriate datasets to evaluate offensiveness score. Therefore, we construct new datasets for the abusive language detection and sentiment classification and utilize them for evaluating our systems for the offensiveness analysis. To handle Korean data, we utilize a refined morpheme-level tokenization method, which has an effect of data augmentation and subword regularization. We summarize our contributions as follows.

- We present a novel demonstration system that analyzes the offensiveness intensity of Korean text based on the abusive language and sentiment information.

- We construct and publicly release a novel dataset for abusive language detection and the sentiment analysis of Korean text.

- Our experiments demonstrate that the multi-task learning of the abusive language detection and the sentiment analysis helps improve the performance of both tasks.

2 System Design

The architecture of KOAS is shown in Figure 2. The system mainly consists of two parts: a classification model and an offensiveness scoring function. We will describe them in detail in the following subsection.

Model Architecture The overall model architecture was inspired by the gated multi-task learning framework (Kim et al., 2019). We denote the proposed neural network as CNN$_{MTL}$ and compare it with the baseline of the pipelined vanilla CNN (Kim, 2014). CNN$_{MTL}$ jointly learns two text analysis tasks - abusive language detection and sentiment analysis. The model is trained with two tasks, learning both linguistic perspective and semantic perspectives simultaneously. This leads to a more computationally efficient model with better performance than training two separate models for each task. Utilizing task-specific layers as well as a shared layer has proven to be effective in learning not only task-dependent features but also useful common features (Kim et al., 2019; Misra et al., 2016; Liu et al., 2017).

The neural model of KOAS consists of an embedding layer, two task-specific convolution layers for the sentiment analysis and abusive language detection, a shared convolution layer, and two softmax layers.

The embedding layer transforms an arbitrary-length input sentence into a matrix of embeddings, denoted as $S = w_1 \oplus w_2 \oplus ... \oplus w_s$, where $S \in R^{s \times k}$, $\oplus$ denotes the concatenation operator, and s and k are the number of words and the dimension of word embedding, respectively. Then, the embedding matrix is fed into the three convolution layers. The features h_{se}, h_{ab}, h_{sh} are obtained from the convolution layers. All the convolution layers are followed by ReLU, max pooling, and dropout layers. The features are then concatenated and fed into the softmax layers. Additionally, we employ a gate mechanism (Hochreiter and Schmidhuber, 1997; Chung et al., 2014) when incorporating task-private convolution output to the other task. We introduce two gates G_{s2a} and G_{a2s}, which control the flow of features from the sentiment analysis task to the abusive language detection task and vice versa, respectively. The gates share useful features and prevent irrelevant information from being propagated. This is calculated by

$$G_{s2a}(h_{se}) = \sigma(W_{s2a} h_{se} + b_{s2a}), \qquad (1)$$

where W_{s2a} and b_{s2a} denote learnable weights and biases, respectively, and σ represents the sigmoid function. W_{s2a} is trained to borrow private features of sentiment analysis task taking h_{se}, which is the output of convolutional layer for sentiment analysis.

$$y^{ab} = Linear(h_{ab} \oplus h_{sh} \oplus G_{s2a}(h_{se})), \qquad (2)$$

$$\hat{y^{ab}} = softmax(y^{ab}) \qquad (3)$$

where h_{ab} and h_{sh} denote output of the convolutional layer for the abusive language detection and the shared convolutional layer for the both task respectively, and $\oplus$ denotes the vector concatenation. All features are concatenated as the input of the fully connected layer. The output for abusive language detection y^{ab} can be derived by Eq. 2. Thereafter, the probability of abusive detection $\hat{y^{ab}}$ is calculated using the softmax function, as shown in Eq. 3. The same method was applied to the sentiment analysis task.

Text Offensiveness The score for text offensiveness is based on the following rules:

- The score of offensiveness represents the degree of negative feelings (e.g. anger, annoyance and fear) that the text may arouse in readers.
- Some abusive expressions may be used as an emphasis or exclamation whereas most abuse expression itself can arouse displeasure regardless of the context.

Therefore, negative tones and abusive expression in the text have a positive correlation with offensiveness, whereas a positive tone has a negative correlation. Based on these hypotheses, we propose a new method for the quantification of offensiveness:

$$O = \sigma(\alpha \times (y^{neg} - max(0, y^{pos})) + \beta \times y^{ab}) \quad (4)$$

where y^{neg} and y^{pos} represent the output values for *negative* and *positive* input sentences, respectively. α and β are hyperparameters that determine the weights of sentiment polarity and abusive expression, respectively and we empirically set α to be 0.456 and β to 0.758. [3]

The proposed method uses y^{neg} and y^{pos}, and the model's prediction of negative and positive polarity. We prevent y^{pos} to be negative, we limit minimum value to zero by applying max function. We use the output value. y^{ab} and y^{neg} represent the degree of how explicitly a profanity appears and how intensely the negative sentiment is expressed, respectively. We have empirically verified that the score becomes correlated with the human feedback of offensiveness. Our experiments demonstrate the practicality of the score when applied to real-world data.

3 Evaluation

3.1 Dataset

Data Construction To ensure that KOAS to handles sentences from various domains, we gathered our training data from three different sources, which covers various domains – YouTube[4], Naver Movie review[5] and dcinside[6]. We scraped comments from a popular Korean online community, expecting our train dataset to be close to a raw expression for practicality. Then, we removed duplicate comments and filtered out non-Korean sentences. The collected dataset consists of 46,853 Korean sentences and was labeled by three annotators on predefined criteria for abusive language (Koo and Seo, 2012) and sentiment following the instruction (Nakov et al., 2016).

Sentences are categorized into binary classes whether they contain abusive language and three classes for sentiment polarity: positive, neutral and negative. Table 1 shows the composition of the classes and the definition of data distribution. The dataset could be downloaded from the link[7]. We hope our corpus can be used for analysis and modeling on Korean abusiveness expression.

Abusive Language		Sentiment		
Abusive	Non-abusive	Pos.	Neu.	Neg.
27%	73%	13%	63%	24%

Table 1: Statistics of dataset label for each task. Pos., Neu. and Neg. indicate positive, neutral and negative polarities respectively.

Sentence Original	볼수록 재수없네 (The more I see him, the worse I get him.)
Morpheme (Mecab)	볼수록 재수없 네
Morpheme (Komoran)	보 ㄹ수록 재수 없 네

Table 2: Original sentence in train set is augmented by two different tokenizers, Mecab and Komoran, based on different parsing rules.

Preprocessing Korean is an agglutinative language (Song, 2006) in which words are constructed with an agglutination of morphemes, and has a syntactic structure different from English. Thus, jamo-level (Stratos, 2017) or morpheme-level tokenization rather than simple word-level tokenization, has been used on the Korean dataset (Park et al., 2018). We applied morpheme-aware subword tokenization, which has proven to be the best tokenization method for Koreans (Park et al., 2020b). To tokenize words at the morpheme level, we utilize KoNLPy, an open-source library for Korean text that provides a number of different tokenizers with different parsing rules and methods. In the training process, we augmented two types of tokenized sentences from each sentence in Korean text with two different tokenizers, Mecab and Komoran, as illustrated in Table 2. Not only does it augment the size of training data around

[3] We empirically set initial α and β to 0.5 and 0.8. Then we tuned α and β as trainable parameters using the Korean Toxic speech corpus (Moon et al., 2020b) with labeled toxic sentences.

[4] https://youtube.com

[5] https://github.com/e9t/nsmc

[6] https://www.dcinside.com/

[7] https://drive.google.com/file/d/1YZ_tuJzs5CBaO0pNY7Cb1Xa0rRR3GQX-/view?usp=sharing

	Positive		Negative	
	Comment	Offensiveness score	Comment	Offensiveness score
Profanity	**존나** 행복하다 **시발** (I'm fucking happy)	3.71	**지랄**이야 **시발** (Fucking hell)	51.89
	미친 존나 재밌네 (It's crazy funny)	19.98	**시발** 나가죽어라 (Fuck off and die)	71.70
	개 예쁘다 (Fucking pretty)	38.37	**게이년** (Gay bitch)	97.78
No Profanity	밝은 것만 보자, 우리 (Let's look on the bright side)	4.06	갈비뼈 순서 뒤집히고 싶어? (Do you want to reverse the order of your ribs?)	56.59
	무대 찢었다 (That was awesome)	19.48	얼굴 못생겼어 (You look ugly)	72.36
	걱정하지마 다 잘될거야 (Don't worry, it'll be fine)	36.36	찐따가 좋댄다 (Do you like it? looser)	81.60

Table 3: Qualitative examples about four combinations, positive-negative and profanity-no profanity. The profanity words are in bold.

Test A@1	$Data_{org}$		$Data_{aug}$	
	AD	SC	AD	SC
Baseline				
CNN (Kim, 2014)	90.52	73.36	90.61	77.54
Ours				
CNN$_{MTL}$	90.82	**80.21**	91.24	79.02
CNN$_{MTL}$ w/o G_{s2a}	**90.92**	79.95	**91.79**	**79.92**
CNN$_{MTL}$ w/o G_{a2s}	90.81	79.81	91.33	78.68
CNN$_{MTL}$ w/o G	89.37	77.03	90.64	79.14

Table 4: Model performance on each setting of gate mechanism with the original data, $Data_{org}$ and with the augmented and over-sampled data $Data_{aug}$. "w/o" and "w/o G" represents "without" and "without G_{s2a} and G_{a2s}". AD denotes model's performance in abusive language detection task, and SC in sentiment analysis. The best results are in bold.

10% on average, but also has the effect of subword regularization (Kudo, 2018; Park et al., 2020a). Accordingly, our model utilizes various sets of subtoken candidates, that yield robustness to typos or slangs.

3.2 Experimental Settings

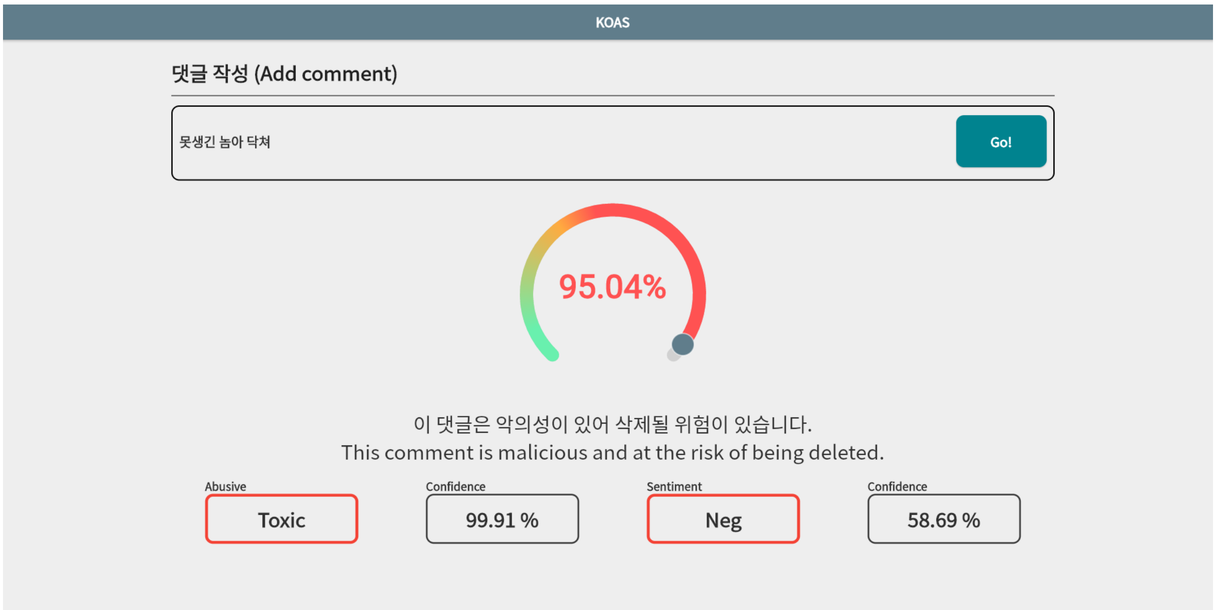
We first split the dataset into a train set (28,111), validation set (9,370), and test set (9,370). To alleviate the class imbalance problem in the train dataset, we over-sampled the minority class dependeding on the dataset size and class proportions (Chawla et al., 2002). We experiment with over-sampling with two different insufficient labels, non-abusive and negative.

3.3 Results

Internal Tasks Table 4 presents the performance of our model on two internal tasks. We test the model performance on the variants of gate mechanism and with different preprocessing steps. There are three main findings: (1) Multi-task learning framework between two tasks generally improves the performance of the model. We observe that CNN$_{MTL}$ obtain higher test accuracy than vanilla CNN's, especially in sentiment analysis. (2) Among four types of multi-task learning architecture, CNN$_{MTL}$ without G_{s2a} is found to be the most effective. Additionally, we empirically validate that the augmented data with over-sampling on non-abusive labels works well according to the $Data_{aug}$ results. (3) With the model accuracy over 90% in abusive language detection and over 80% in sentiment analysis, KOAS has the potential of being competently extended to the other downstream applications where a detailed analysis on offensiveness is required.

Text Offensiveness To validate the legitimacy of the computed text offensiveness score, we measured the Pearson correlation between the predicted score and human feedback on offensiveness. We randomly chose 300 sentences from the test set, and labeled each sentence in three classes, regardless of whether the sentence is not offensive, mildly offensive or strongly offensive. The score from the model's prediction has a Pearson correlation coefficient of 0.62, which implies that the score has a positive correlation with human feedback.

Qualitative Examples To test our system KOAS, we classify six types of real-life sentences: positive-



댓글 작성 (Add comment)

못생긴 놈아 닥쳐

이 댓글은 악의성이 있어 삭제될 위험이 있습니다.
This comment is malicious and at the risk of being deleted.

Abusive	Confidence	Sentiment	Confidence
Toxic	99.91 %	Neg	58.69 %

Figure 3: KOAS user interface : the system receives “못생긴 놈아 닥쳐 (Shut up, ugly bastard)” as input. Offensiveness score is nearly 95% with abusive and negative prediction.

abusive, neutral-abusive, negative-abusive, positive-non abusive, neutral-non abusive, negative-non abusive. Table 3 presents some qualitative examples with offensiveness score about four types. In general, sentences containing swear words have higher offensiveness score by the abusive detection. We see that our system reflects sentiment polarity, we observe different offensiveness scores, even they exhibit similar profanity such as “존나 행복하다 시발”-“I’m so fucking happy ” and “시발 나가죽어라”-“Fuck off and die”, one is 3.71, and the other is 71.70 in Table 3. It relies on words that appear together. We test various negative and non-abusive expressions even they don’t contain swear words. it tends to detect implicit negative tone, such as “얼굴 못생겼어”-“Your face is ugly”, which obtained 72.36, compared to positive and non-abusive examples such as “밝은 것만 보자, 우리”-“Let’s look on the bright side”, which obtained 4.06. On the other hands, some examples with various negative tones (e.g. sarcastic tone) show still challenging to detect such as “성형많이 해서 존나 예쁘다’-“She’s so pretty with a lot of plastic surgery".

4 Demonstration

KOAS has an intuitive and simplified interface, where users can type any Korean sentence and check the offensiveness of the input sentence. When KOAS receives a sentence, it internally conducts abusive language detection and sentiment analysis, and then computes the offensiveness score. The logit value of the sentence containing an abusive expression and the most likely sentiment polarity is shown at the bottom, as well as the offensiveness score. Figure 3 shows a user’s input interface of KOAS. There is a status bar in the middle of the screen, so users can check the level of offensive intensity. When the score is higher than a predefined *moderate* threshold, messages like “This comment might have malicious intent.” and *crucial* threshold, warning messages like “This comment is malicious and at the risk of being deleted.” appear on the screen. In this work, we set the moderate threshold to be 40 and crucial threshold to be 72.

5 Conclusion

In this paper, we have proposed KOAS, a novel system that efficiently estimates the offensiveness score of Korean text. We expect KOAS to attenuate the usage of offensive languages by providing a real-time feedback to writers about their writing. KOAS has notable technical novelty and social contributions, including (1) tackling morphological richness and complex syntax of Korean, (2) incorporating linguistic and contextual analysis to capture the offensive nuance of text and (3) effectively analyzing the offensiveness of text. Our CNN-based system is lightweight, practical and applicable to various hardware environments compared to transformer-based system.

Some usecases we expect are as follows: First, for the people unfamiliar with Korean, the KOAS system can be used to prevent unintended attacks when they post Korean articles and help them recognize attacks in Korean sites. Second, site administrators who need to block offensive comments can port the KOAS system to automatically block comments that exceed a certain score. Finally, our datasets can be utilized for further research on Korean text analysis, including Korean language understanding and automatic labeling. We expect that our proposed system will be readily applicable to various downstream applications, including education, game, real-time chatting and social media platforms.

6 Future Work

Following experiments and qualitative examples, we have found that real-world sentences containing various negative tones without abuses are still challenging because of their implicit offensiveness. We plan to build our system on recent language models such as KoBERT[8] and KoELECTRA[9], which is expected to make our system highly reliable and robust.

7 Acknowledgements

We thank the anonymous reviewers for their helpful comments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2021R1A2C3010430) and the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A4A1018309).

References

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Timothy Jay. 2009. Do offensive words harm people? *Psychology, public policy, and law*.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada. Association for Computational Linguistics.

Kang-Min Kim, Yeachan Kim, Jungho Lee, Ji-Min Lee, and SangKeun Lee. 2019. From small-scale to large-scale text classification. In *The World Wide Web Conference*.

Kwang-Young Kim, Seo-Young Jeong, Jung-Hoon Park, Seok-Hyoung Lee, Hye-Jin Lee, Jae-Wook Seol, Chul-Su Lim, and Jung-Sun Yoon. 2018. Performance comparison of korean keyword-based document classifiers using convolutional neural networks. *International Journal of Applied Engineering Research*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

H Koo and E Seo. 2012. A study on injurious comment spam-its typology and suggestions for improvement. *Korean Language Research, null (30)*, pages 5–32.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Rebecca S Merkin. 2009. Cross-cultural communication patterns-korean and american communication. *Journal of intercultural communication*, 20.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Tackling online abuse: A survey of automated abuse detection methods. *arXiv preprint arXiv:1908.06024*.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020a. Beep! korean corpus of online news comments for toxic speech detection. *arXiv preprint arXiv:2005.12503*.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020b. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Jungsoo Park, Mujeen Sung, Jinhyuk Lee, and Jaewoo Kang. 2020a. Adversarial subword regularization for robust neural machine translation. *arXiv preprint arXiv:2004.14109*.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020b. An empirical study of tokenization strategies for various korean nlp tasks. *arXiv preprint arXiv:2010.02534*.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. 2018. Subword-level word vector representations for korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[8] https://github.com/SKTBrain/KoBERT

[9] https://github.com/monologg/KoELECTRA

Peyman Passban, Qun Liu, and Andy Way. 2018. Improving character-based decoding using target-side morphological information for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.

Jae Jung Song. 2006. *The Korean language: Structure, use and context*. Routledge.

Karl Stratos. 2017. A sub-character architecture for korean language processing. *arXiv preprint arXiv:1707.06341*.